PUC RIO

Dissertação de Mestrado

LLMs as reflective information retrieval partners for academic knowledge work: A generative probing approach

João Henrique Gallas Brasil

Pontifícia Universidade Católica do Rio de Janeiro
Centro Técnico Científico
Departamento de Informática

Rio de Janeiro, 24 de Setembro de 2025

Pontifícia
Universidade
Católica do
Rio de Janeiro

Dissertação de Mestrado

LLMs as reflective information retrieval partners for academic knowledge work: A generative probing approach

João Henrique Gallas Brasil

Orientação: Profa. Simone Diniz Junqueira Barbosa

Dissertation presented to the Programa de Pós–graduação em Informática, do Departamento de Informática of PUC-Rio, in partial fulfillment of the requirements for the degree of Mestre em Informática.

Rio de Janeiro, 24 de Setembro de 2025

Pontifícia
Universidade
Católica do
Rio de Janeiro

LLMs as reflective information retrieval partners for academic knowledge work: A generative probing approach

João Henrique Gallas Brasil

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Informática. Aprovada pela Comissão examinadora abaixo:

Profa. Simone Diniz Junqueira Barbosa
Orientadora
Departamento de Informática – PUC-Rio

Prof. Helio Côrtes Vieira Lopes
Departamento de Informática – PUC-Rio

Prof. Renato Fontoura de Gusmão Cerqueira
Instituto PUC-Behring de Inteligência Artificial – PUC-Rio

Rio de Janeiro, 24 de Setembro de 2025

Pontifícia
Universidade
Católica do
Rio de Janeiro

João Henrique Gallas Brasil

Earned his bachelor's degree in Design from the Universidade do Estado do Rio de Janeiro. Completed his master's degree at the Departamento de Informática da PUC-Rio, specializing in Human-Computer Interaction.

Ficha Catalográfica

Brasil, João Henrique Gallas

LLMs as reflective information retrieval partners for academic knowledge work – a generative probing approach / João Henrique Gallas Brasil; advisor: Simone Diniz Junqueira Barbosa. – 2025.

119 119 f: il. color. ; 30 cm

Dissertação (mestrado) - Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2025.

Inclui bibliografia

Interação Humano-Computador; Grandes Modelos de Linguagem; Trabalho de Conhecimento Acadêmico; Geração Aumentada por Recuperação; Busca de Informação; Ferramentas Epistêmicas; Sistemas de Iniciativa Mista; Busca de Informação Generativa. Diniz Junqueira Barbosa, Simone. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

Acknowledgments

Thanks to Elisa, my parents, sister, and grandmother, and to Andrea, who were, even from a distance, with me throughout this project.

Thanks to Simone, Renato, Hélio, and Alberto.

And to all my colleagues at IBM Research and PUC with whom I spoke, even if for just a few minutes, during the process of this work. Without some of the briefest of these conversations, it would not have been completed.

In summary, to all the giants on whose shoulders I stand.

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Finance Code 001.

Abstract

Brasil, João Henrique Gallas; Diniz Junqueira Barbosa, Simone (Advisor). **LLMs as reflective information retrieval partners for academic knowledge work: a generative probing approach**. Rio de Janeiro, 2025. 119p. Dissertação de Mestrado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

The wide availability of information on the web has intertwined the processes of learning and conducting research with the challenges of finding and interpreting a large volume of possibly relevant content. In this context, Large Language Models (LLMs) have emerged as a potential tool, with retrieval augmented generation (RAG) techniques now enabling them to engage with the information most relevant to the user at interaction time.

However, most LLM-based applications currently prioritize generation over retrieval, resulting in question-and-answer tools not necessarily suited to the academic research workflow. This complex kind of knowledge work requires the coupling of information retrieval and management with the synthesis, contextualization, and critical evaluation of information.

To explore these challenges, we conducted a review of the related work in this space, alongside an interview study involving ten researchers from diverse STEM fields and varying expertise levels. Our aim was to understand their current work practices and perceptions of AI tool use. Based on our findings, which highlight a need for tools that support exploration rather than just providing answers, we propose Generative Retrieval Probing – an interaction format for LLM-based systems that emphasizes user-led discovery and critical engagement with source material. We instantiated this paradigm in a prototype system integrated with a researcher's personal document collection, which we then tested with eight researchers.

Keywords

Human-Computer Interaction; Large Language Models; Academic Knowledge Work; Retrieval-Augmented Generation; Information Seeking; Epistemic Tools; Mixed-Initiative Systems; Generative Information Retrieval.

Resumo

Brasil, João Henrique Gallas; Diniz Junqueira Barbosa, Simone. **LLMs como parceiros reflexivos para recuperação de informação no trabalho acadêmico: uma abordagem de sondagem generativa**. Rio de Janeiro, 2025. 119p. Dissertação de Mestrado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

A ampla disponibilidade de informações na web tornou os processos de aprendizado e pesquisa intrinsecamente ligados aos desafios de encontrar e interpretar um vasto volume de conteúdo potencialmente relevante. Nesse contexto, os Grandes Modelos de Linguagem (LLMs) surgiram como uma ferramenta em potencial, com técnicas de Geração Aumentada por Recuperação (RAG) que lhes permitem recuperar as informações mais relevantes para o usuário no momento da interação.

No entanto, a maioria das aplicações atuais baseadas em LLMs prioriza a geração em detrimento da recuperação, resultando em ferramentas de pergunta e resposta que não são necessariamente adequadas ao fluxo de trabalho da pesquisa acadêmica. Esse tipo complexo de trabalho de conhecimento exige a articulação entre a recuperação e gerenciamento de informações e a síntese, contextualização e avaliação crítica.

Para explorar esses desafios, realizamos uma revisão dos trabalhos relacionados e um estudo de entrevistas com dez pesquisadores de diversas áreas de STEM e com diferentes níveis de experiência. Nosso objetivo foi compreender suas práticas de trabalho atuais e suas percepções sobre o uso de ferramentas de IA. Com base em nossos achados, que evidenciam a necessidade de ferramentas que apoiem a exploração em vez de apenas fornecer respostas, propomos a Sondagem de Recuperação Generativa ("Generative Retrieval Probing") – um formato de interação para sistemas baseados em LLMs que enfatiza a descoberta guiada pelo usuário e o engajamento crítico com as fontes. Instanciamos esse paradigma em um protótipo integrado à coleção pessoal de documentos de pesquisadores, que foi então testado com oito participantes.

Palavras-chave

Interação Humano-Computador; Grandes Modelos de Linguagem; Trabalho de Conhecimento Acadêmico; Geração Aumentada por Recuperação; Busca de Informação; Ferramentas Epistêmicas; Sistemas de Iniciativa Mista; Busca de Informação Generativa.

Table of Contents

List of Figures

List of Abbreviations

- **ML**: Machine Learning
- **API**: Application Programming Interface
- **Gen-IR**: Generative Information Retrieval
- **GRP**: Generative Retrieval Probing
- **LLM**: Large Language Model
- **RAG**: Retrieval Augmented Generation
- **STEM**: Science, Technology, Engineering, and Mathematics
- **UI**: User Interface
- **DIKW**: Data, Information, Knowledge, Wisdom
- **GenRec**: Generative Recommendation
- **GR**: Generative Retrieval
- **RRG**: Reliable Response Generation

1
Introduction

The wide availability of information on the web has intertwined the processes of learning and conducting research with the challenges of finding and interpreting a large volume of possibly relevant content. While Large language models (LLMs), augmented with retrieval mechanisms, have emerged as powerful tools for navigating this landscape, their common implementation as answer-providers often fails to support the reflective and critical nature of academic knowledge work. This dissertation explores how LLM-based systems can be designed to function as reflective epistemic tools that aid researchers in interpreting, synthesizing, and building upon their own work, rather than simply delivering definitive answers.

To ground this exploration, we begin in Section 2 by reviewing the background literature on academic workflows, epistemic practice tools, and the technical foundations of LLMs and retrieval-augmented generation. Section 3 then situates our contribution within the related work, examining the nascent field of generative information retrieval (GenIR) and existing applications of LLMs for academic information seeking.

Section 4 details a preliminary qualitative study with STEM researchers to understand their current information management practices and perceptions of AI tools. Based on these findings, in Section 5 we propose and detail the architecture of Generative Retrieval Probing (GRP), a novel interaction format and prototype system designed to support user-led discovery through proactive and reactive engagement with a researcher's personal literature collection.

The evaluation of this prototype is the focus of Section 6, which presents the results of a second qualitative study where researchers interacted with the GRP system using their own curated document collections. Finally, Section 7 synthesizes the findings from our studies to offer broader conclusions on designing LLMs as reflective tools and outlines promising directions for future work in this domain.

1.1 Motivation

The modern academic landscape, marked by the vastness of digitally available works, presents a dual challenge to researchers: they must not only generate new knowledge, but first navigate and synthesize an overwhelming volume of existing content. This reality has fundamentally altered the nature of academic work, transforming it into a process where effective information seeking and sense-making are as crucial as the research itself.

In parallel, and as a consequence of this vast accessibility, large machine learning (ML) models have emerged. Specifically, large language models (LLMs), trained on massive corpora of internet content such as CommonCrawl and WebText2 (Brown et al., 2020), are capable of generating human-like text based on written prompts. In basic LLM systems, this information, consisting of reproducing patterns from internet data, is ingrained through training techniques that convert semantic structures (broadly, words) into numerical tokens.

Building on this foundation, retrieval-augmented generation (RAG) (Lewis et al., 2021) and related methods (Wang et al., 2024a), (Gao et al., 2022) enable LLMs to incorporate new tokens during text generation in a "zero-shot" manner, meaning without additional training. These tokens are typically retrieved from source texts using semantic similarity techniques to ensure the most relevant information is used based on the input. In theory, this allows LLM-based systems to engage with new information in real time.

This forms the basis of many systems that purport to use LLMs as tools to process, manage, or even create "knowledge". Examples include PDF processing tools like SciSpace [1] and Microsoft Copilot [2], as well as academic workflow tools like Elicit [3] and Jenni AI [4]. The chat-based tools that initiated this wave, such as ChatGPT and Anthropic's Claude, are also incorporating features like multimodal file uploads and persistent memory of user interactions, enabling users to manage their information across multiple conversations.

1.2 Problem definition

The popularity of these LLM+RAG-based tools has given rise to concern that, due to their generative nature and the general-purpose origins of their

[1] https://scispace.com/ - All URLs accessed on 28 July 2025 unless specified otherwise.

[2] https://copilot.microsoft.com/onboarding

[3] https://elicit.com/

[4] https://jenni.ai/

underlying LLMs, they might be thinking for users rather than truly assisting them in knowledge work. The impact of LLMs on cognition and analytical thinking is a topic of growing academic concern. Prominent recent work by Kosmyna et al. (2025), for instance, provides compelling neurological and behavioral evidence that reliance on LLM assistants for complex writing tasks can diminish critical engagement and lead to a phenomenon they term cognitive debt. Their findings, pointing towards LLM use correlating with weaker neural connectivity and impaired memory encoding, are strongly related to the design concern motivating this dissertation.

A core aspect of this recent study is its direct contrasting of the cognitive states induced by LLM use versus those from traditional search engines. They conclude that while search engines prompt active information seeking, the synthesized, singular responses from LLMs encourage a shift toward passive consumption. This dichotomy between active, user-led retrieval and passive, system-led generation is a foundational theme we also explore, albeit from a design and systems perspective.

While this work proceeds with its own investigation into a proposed solution, we will return to the important implications of the work by Kosmyna et al. in our concluding chapter. There, we will more deeply address how our proposed Generative Retrieval Probing framework relates to their findings and to the fundamental dichotomy they highlight between different modes of technology-assisted knowledge work.

In this direction, a core reflection of this work is that the broadening use of LLMs for knowledge work tasks brings them into the realm of knowledge and epistemic practices, where a well-established domain of tools exist which focus far less on quick access to all-encompassing answers. Instead, they focus on serving as external representations of knowledge that bridge the material and symbolic aspects of practice, helping users to manipulate, interpret, and generate new knowledge (Markauskaite and Goodyear, 2017).

Additionally, for academic workflows, there is also a variety of tools currently in use that leverage non-LLM machine learning algorithms in various ways. One prominent example is that of academic paper recommendation systems, which use algorithms to suggest relevant research. While these tools vary in their effectiveness and adoption, they represent a type of AI that researchers commonly encounter in literature search systems and academic databases (Beel et al., 2016).

From this outlook, we propose to study how LLMs might interface productively with the patterns and formats found in these types of tools. In this work, we consider that RAG-enabled LLMs are not in fact sufficient as standalone knowledge tools, but can be components of productive knowledge retrieval and management systems. We seek to explore this in the domain of academic research, with a particular focus on literature-seeking workflows, and literature and knowledge management tools. In essence, our guiding research question is:

How can LLM-based systems be designed to function as reflective epistemic tools for academic research, shifting their role from providing definitive answers to helping researchers interpret, synthesize, and build upon their own work?

From this core question, we draw the further subquestions, which were the ones that guided our process throughout this work:

1. What are researchers' current workflows for literature knowledge, and what are the key challenges and opportunities for AI in that environment? (Addressed in Section 4)

2. What design principles and interaction format can position an LLM as a reflective tool, shifting its outputs from full answers to outputs that invite further inquiry? (Addressed chiefly in Section 5)

3. How do researchers interact with and perceive a reflective AI tool, and what is the effect of this interaction on their cognitive processes and established workflows? (Addressed in Section 6 and Section 7)

1.3 Methodology

Drawing from these questions, this research employed a multi-stage qualitative design to first understand the problem space of academic knowledge work and then design and evaluate a technological intervention. The process was divided into three primary phases: an initial exploratory study to identify key challenges and opportunities; the design and implementation of a prototype system, *Generative Retrieval Probing*; and a final evaluative study to assess the prototype's utility and alignment with research workflows.

First, to understand how researchers currently engage with academic literature and technology, we conducted an exploratory qualitative study. This investigation was guided by the need to ground our work in the authentic practices and perceptions of our target users. We conducted semi-structured interviews with a diverse sample of 10 STEM researchers, spanning different career stages and both academic and industrial research contexts. The interviews explored their information-seeking and management behaviors, organizational strategies, their use of recommendation and AI-based tools, and how they perceive connections within the literature they are working with. The collected data was transcribed and analyzed using thematic analysis to identify the core patterns and needs that would inform the subsequent design phase.

Building on the insights from this initial study, we proceeded to the design and implementation of a functional prototype system realizing an interaction format we term Generative Retrieval Probing (GRP). This system embodies our proposed interaction format for an LLM-based tool designed to support the reflective and exploratory nature of academic research. Composed of a local, self-hosted backend and a browser extension that integrates with the Zotero reference manager, GRP uses a retrieval-augmented generation (RAG) engine to process not only a researcher's literature collection but also their explicit interactions with it, such as notes and highlights, to generate verifiable, context-aware suggestions intended to prompt reflection rather than provide definitive answers.

Finally, we conducted a second qualitative study to evaluate the GRP prototype. This evaluation aimed to investigate how researchers would interact with the tool and perceive its utility. The study employed a two-part think-aloud protocol with eight researchers. In an initial preparatory session, participants curated personalized document collections drawn from their own research, ensuring the inclusion of topics with which they were both familiar and unfamiliar. In the second session, they engaged with the GRP prototype using their own collections, exploring its suggestions, using its features to generate new insights based on their own actions, and narrating their thought processes, in a think-aloud format. These sessions were screen- and audio-recorded, and the resulting data was transcribed and thematically analyzed to understand how researchers perceived the tool, its AI-generated outputs, and its potential fit within their established research practices.

Throughout all phases, a continuous review of related literature on generative information retrieval and academic information seeking was conducted. This served to contextualize our findings, identify gaps in existing research, inform the theoretical foundations and design principles of the GRP system, and position our contribution to the field.

1.4 Contributions

We anticipate that this research is meaningful not just as headway in the creation of a usable experimental tool, but that it will also contribute to the broader discourse on integrating Large language models into knowledge practices. The primary contribution of this work is the proposal and initial validation of Generative Retrieval Probing, an interaction format that embodies a different design philosophy for these tools. It seeks to shift the role of LLMs away from being providers of definitive answers and toward being reflective tools that support the reflective and exploratory nature of academic work, addressing a core misalignment with research workflows observed in current systems.

By building and evaluating this experimental framework, this work provides valuable empirical insights into how such a system can support a researcher's cognitive processes. Our findings suggest that this approach can aid in both validating a researcher's existing understanding and in sparking novel connections they had not previously considered. In this way, the GRP format and the insights gained from its evaluation serve as a foundation for future research in this area. We also believe these principles and findings are applicable beyond the domain of academic work, offering a transferable framework for employing LLMs in various knowledge-intensive fields.

2
Background

In the following section, we review some of the fundamental concepts and related work relevant to the design of LLM-based systems for academic knowledge work, detailing our particular focus on academic literature workflows. We begin by exploring the nature of research work, emphasizing its connection to information seeking and knowledge management. We then delve into the characteristics of tools that support knowledge work, highlighting their role as epistemic artifacts that bridge the gap between the material and symbolic aspects of practice. Finally, we shift our focus to large language models (LLMs) themselves, discussing their capabilities, limitations, and the crucial role of retrieval mechanisms in augmenting their functionality for many tasks.

2.1
Academic literature workflow, information seeking, and information management

The concept of academic research work is difficult to define precisely due to the multitude of forms it can take and the variety of research philosophies that have been prevalent throughout history (Post-positivism, Constructivism, Constructionism, among others (Tamminen and Poucher, 2008)). Nichol et al. Nichol et al. (2023) offer the broad perspective that academic research entails developing a complex of interrelated concepts and arguments, including solutions to a range of meta-questions, implicitly for the common goal of then applying these concepts and arguments into a coherent written work for publication.

The workflow that follows from this process, then, consists of the activities that support the development of these concepts and arguments, commonly by drawing on information from existing academic literature. This connects academic work intimately to the broader field of information seeking, and due to the challenges of storing, organizing, and retrieving or re-finding information when relevant also brings a clear connection between academic research and the field of information management. These particular activities, and their implications for the subsequent writing of academic papers, define a more specific academic literature workflow, which serves as a particular point of focus for this work.

Modern academic information seeking is predominantly conducted through search engines and digital libraries, which have become the default starting points for research (Hoeber and Storie, 2022). These are the environments in which a great many modern research strategies actually occur, *de facto*, with researchers translating their overall information seeking goals into interactions with these tools. Consequently, researchers adopt a range of noteworthy behaviors that go beyond simple lookup-based retrieval and are closely tied to broader information-seeking needs.

A dynamic search pattern involving iterative query building (also called berrypicking) is the most widespread of these behaviors, where the results of each lookup bring insights that change the search query being used. Beyond queries, it is also common to perform research in which iteration contributes generally to a better understanding of a domain through multiple stages or more exploratory intermediary search steps, termed multi-stage and exploratory search, respectively (Vakkari, 2001). Academic information seeking, in particular, includes a crucial component of filtering and verifying information as it is found, which plays into both of these forms of iterative seeking. In some cases, these behaviors and broader research strategies around them are about extending or updating existing knowledge for a particular task, while in others they might be about building an initial understanding of a completely new domain. Moreover, the citation-driven nature of academic texts makes citation-based seeking, such as forward and backward chaining, a fundamental secondary information seeking strategy, which can be present at any moment in conjunction with other patterns (Hoeber et al., 2019). While these are useful general descriptions of behaviors, research is a broad practice performed on many different levels, from undergraduate students to professional post-doctoral researchers, and actual practices vary not only with the complexity of the task at hand and the specific research topics, but with the level of experience of the individual.

Likewise, modern information management for academic work is most frequently supported by digital tools, with most researchers working with files only in lieu of printed papers (Osae Otopah and Dadzie, 2013). Management of these entails working or engaging with information beyond reading, a sometimes implicit activity where researchers and students interpret and make sense of academic information. They identify core themes, relevant insights, and valuable details from a larger body of literature. This involves developing ways to decide what constitutes valuable information for them, personally.

This information management can take many forms, from clustering in organization by the overall recognition of patterns across information being worked with, to taking notes, a form of interaction with academic information to make it more manageable, meaningful, and interpretable (Conrad et al., 2020). Dealing with information sources themselves, such as the aforementioned libraries, is also a factor. Ultimately, many aspects of information management are expressions of organization. Behaviors such as piling and filing describe how researchers manage large volumes of literature, either by grouping related items or creating an organized structure to store them (Mizrachi and Bates, 2013). In addition, personal goals and current objectives have an impact on organization, as they are subjective and vary from person to person, but are ultimately connected to personal needs and context. The use of citation management software is widespread, though not universal (as we later saw in our sample of researchers for the first study, in section 4), in organizing research, as the need to include referenced information in written work is greatly facilitated by the ability to manage metadata, compared to simply maintaining a collection of files (Kathleen Kern and Hensley, 2011).

2.2
Information, knowledge, work, and epistemic practice tools

The aforementioned development of a coherent complex of interrelated concepts and arguments for one's use, in this case drawing from information that can be found in academic literature, is one example of what situates academic work as knowledge work. While precise origins are unclear, there is a long established hierarchical relation between information and knowledge in literature, such as in the DIKW model, which also incorporates data (below information), and wisdom (above knowledge) (Zeleny, 1987). The distance between them is where interpretation, application, and situation in a particular domain of use happens. This distinction highlights the cognitive processes required for meaningful knowledge acquisition, which involve not just the retrieval, but the critical evaluation, synthesis, and contextualization of information.

This view of knowledge work, and academic work alongside it, is well situated in the broader context of epistemic practice (Cetina, 1991), the activities of inquiry in which it is necessary to confront non-routine problems and devise non-routine solutions by drawing from a significant knowledge base. The definition of epistemic artifacts and epistemic tools are proposed from this, those being artifacts and tools that facilitate knowledge production by bridging the physical and mental aspects of practice, integrating material and symbolic elements to support inquiry, learning, and decision-making. In this context, the distinction between artifact and tool is blurred, and in a digital environment it becomes practically non-existent. A database, for example, can serve as a tool for inquiry, used to manipulate data, but it can also be viewed as an epistemic artifact that encapsulates prior knowledge and organization of information for future use (Markauskaite and Goodyear, 2017).

The importance these tools (or artifacts) have for knowledge work is that they allow users to manipulate these external environments, such as information stores, while engaging in reflective thinking, which allows professionals to both interact with and interpret the world, helping them generate new knowledge (Markauskaite and Goodyear, 2017). Many of the software tools commonly used in academic literature work, as presented above, fit the definition of epistemic tools – search engines, digital libraries, and paper files or reference data serve not only to provide representations of information but also as scaffolds that enable researchers to discover, refine, and expand upon existing knowledge, and through this fuse manipulation of information with symbolic understanding.

Personal knowledge management tools, which see some use among researchers, are a noteworthy additional example, such as note-taking and note-repository tools influenced by the zettelkasten method (Luhmann, 1981). These are focused on helping users manage, retrieve, and connect information through a persistent external representation of their knowledge, which is most often achieved using secondary artifacts, such as backlink graphs, to create and visualize these connections (Girard and Girard, 2015).

Drawing from the epistemic dimension of tools, this object of reflective thinking was a particular focus to us in this work throughout both the initial investigations and later evaluation, as a more general description of how researchers must actively engage with information to transform it into knowledge. Rather than passively receiving information, the process of reflection involves questioning, comparing, and synthesizing information from various sources to build a coherent understanding. This is particularly relevant in the context of academic work, which is not merely about finding correct answers, but about developing a nuanced and critical perspective on a body of literature. The tools that support this work, therefore, should not aim to replace this reflective process, but to provide the necessary structure and material for it to occur more effectively.

2.3 Large language models, retrieval augmented generation, and Information retrieval

Large language models arose as a form of natural language processing system, based on techniques for obtaining vector representations of words through machine-learning model training. This approach has been widely shown to capture the semantic relevance of words in numerical space (Mikolov et al., 2013), crucial to the core idea of language modeling. Further proposals in the realm of attention-based sequence-to-sequence neural networks then led to breakthroughs in natural language generation from natural language inputs (Vaswani et al., 2023). Pre-training on large unlabelled corpora of text sourced from the internet have enabled language models to generate meaningful text from nearly any input (Radford et al., 2018), enabling them to be deployed for general use. This has lead to a boom in language model use, in particular with models tuned to simulate a human-like chat interaction, and deployed in systems resembling chat interfaces that have come to satisfy general information seeking needs for many users (Caramancion, 2024).

A series of subsequent works dealing with generalizations of their capabilities has greatly expanded the employment of large language models in various language-related tasks, including replicating language outputs of human-like reasoning (Wei et al., 2023). These works have demonstrated how the general capability of language generation can be a gateway to many other abilities (Brown et al., 2020). Large Language Models have thus come to be commonly seen as the hallmark "foundation models", meaning ML models capable of being fine-tuned to complete, or to aid humans in completing, a wide variety of tasks, be they ones directly involving language, or beyond (Bommasani et al., 2022).

Systems employing LLMs, such as the aforementioned chat interfaces, are now also increasingly used for knowledge intensive tasks in a variety of domains (Ritala et al., 2024). In this context, a significant limitation of these large language models lies in their pre-trained nature. The information, or "LLM knowledge", they have access to is determined by this training – which is very costly – and thus it is impractical for models to be re-trained to add new information. This means that while LLMs might be able to provide some aid in information seeking for knowledge intensive tasks, they are limited in their answers by the information that was available online at the time of data-collection, and the generative process atop it afforded by the learning.

However, this is a limitation that can be greatly relieved by methods for retrieving context tokens from outside the parameters of a language model, allowing any text to be tokenized and used as a form of non-parametric memory while an LLM is running (Lewis et al., 2021), thus increasing this LLM knowledge using external resources, often ones supplied by users or from application and context-specific databases. This is what is meant by Retrieval Augmented Generation (RAG), which has itself spawned a series of related techniques around this general idea of retrieval. A large part of the more advanced functionalities of common LLM applications (as in the file and memory capabilities of ChatGPT, Claude and Copilot), or for some the entire foundation of their inner workings (as in JenniAI, and Elicit, dedicated to academic), comes from these retrieval approaches, which play into the general view of LLMs as being capable of openly aiding in knowledge-intensive tasks. While what is meant by knowledge in the context of LLM retrieval does not align with definitions of human knowledge, these methods do allow for a great expansion of an LLM's capabilities in providing relevant information.

What this reliance on retrieval mechanisms does, in fact, is transform LLM systems at least partly into information retrieval systems. As a field in computer science, information retrieval is focused on meeting an information need by efficiently tracing and recovering relevant information from data collections, often large and unstructured. However, this strong IR aspect of current LLM systems is often overlooked in favor of the more prevalent chat-based interaction format, which presents LLMs as providers of all encompassing answers. Designing LLM-based tools with a focus on retrieval, or even centering the tool entirely around it, is an as of yet uncommon approach, with only a few outliers that emphasize the retrieval component as the main focus of interaction (Elicit, among the examples given in this section, being an example). In fact, there are already many other productive intersections between academic information seeking and the field of information retrieval, with search engines being the most direct example, and literature recommendation algorithms being another noteworthy application mentioned throughout this work.

2.4 Considerations

While the notion academic research work is broad and difficult to define, for the context of this work we believe focusing on academic information seeking and information management was a productive lens. The retrieval of valuable information, given the current goals of the researcher, is a critical component of both these processes and serves as a bridge between accessing information and integrating it into research workflows.

As outlined above, an important characteristic of existing tools that support knowledge work in these areas is their function as epistemic artifacts. These tools bridge the physical and mental aspects of academic practice, facilitating reflective thinking and enabling researchers to engage with information in ways that support knowledge generation (Markauskaite and Goodyear, 2017). Large language models (LLMs) also emerge as a potentially powerful type of tool, now strongly relying on robust information retrieval mechanisms. However, in current implementations, these mechanisms are frequently underemphasized, overshadowed by chat-based interaction paradigms that abstract away their retrieval components.

By examining how researchers employ information seeking and management tools as epistemic tools and analyzing how these behaviors relate to their practices in information retrieval, we can seek to identify opportunities to more effectively integrate these dynamics into LLM-based information retrieval systems. We believe a deeper understanding of the interplay between these tools and researchers' workflows may guide the development of LLM systems that prioritize and enhance retrieval functionalities, ultimately aligning more closely with these aspects of academic knowledge work.

3
Related work

There are many works that propose and optimize general LLM RAG chat systems for a variety of knowledge intensive domains. Examples of this include medical education (Al Ghadban et al., 2023), cybersecurity evaluation (Ferrag et al., 2024), and the financial sector (Zhao et al., 2024). However, as we argue above, this does not necessarily make these knowledge tools effective in supporting genuine understanding or critical engagement with the content. While these systems can facilitate information access, they do not necessarily address the deeper cognitive processes required for meaningful knowledge work, where the goal is not just to find an answer, but to synthesize, evaluate, and build upon information within a specific context.

To situate our own contribution, this section reviews the literature most relevant to designing LLM-based systems that more directly support these knowledge processes, in the molds described above. We focus on two key areas. First, we will review the nascent field of *generative information retrieval* (Gen-IR), which formalizes studies on LLM-based systems with a stronger focus on the retrieval component. We explore approaches from this field and relate them to the core aspects of our work, both theoretical and technical aspects we will discuss afterwards. We then give a broad review of works on LLMs for academic information seeking, detailing the implications from that area we believe are most relevant to our work.

3.1
Works on LLMs focusing on information retrieval for users

To situate our work, this section reviews the nascent field of *generative information retrieval* (Gen-IR), which formalizes studies on LLM-based systems with a stronger focus on the retrieval component. We begin by establishing the key distinction between two main paradigms: works on *Generative Retrieval* (GR), where models generate document identifiers directly, and those focused on *Reliable Response Generation* (RRG), which augment LLMs with external sources like RAG while prioritizing the retrieval process to ensure verifiability. As our proposal aligns with this second category, the remainder of the section details specific technical approaches within the RRG landscape, beginning with frameworks designed to improve the quality of retrieved context and ensure reliable attribution. We then turn to systems that employ structured, multi-step retrieval through chained queries and graphs of thought, before concluding with a review of various approaches to personalization that leverage both structured user data and sophisticated prompt engineering.

While we established in Section 2.3 and Section 2.4 our particular vision for a possible category of LLM based systems that can be more suitable for information retrieval, a nascent field with this general focus has started to become more present in publications while this work was being written. This area of study has recently reached critical mass. We carried out an initial unstructured literature review for the proposal version of this work in April 2024, where we mapped out eight papers that present LLM-based systems with a stronger focus on information retrieval for users rather than on the generation portion. Since then, a comprehensive literature review by Li et al. Li et al. (2025) has been published, which helps to more firmly establish this nascent field of generative Information Retrieval. Their work and the broader focus for the area that they help to establish resonate with our own concerns regarding the tendency of popular LLM systems to subordinate the retrieval process, treating it as a background mechanism rather than a primary function for serving a user's information-seeking goals. The systematic review, which surveys 161 works, provides a foundational map of current trends and methodologies that we can build upon. Of those, some works perform retrieval by using the LLM to generate document identifiers directly, a practice they term Generative Retrieval, which is also more similar to many works in Generative recommendation. In contrast, some incorporate retrieval into LLM systems by more directly leveraging techniques such as RAG while also prioritizing this retrieval in different ways over most mainstream LLM systems, which Li et al. term Reliable Response Generation.

This second category of systems aligns with our previously detailed views on the potential of information retrieval systems incorporating LLMs. This distinction is also worthwhile, as it separates these works from post-hoc citation generation approaches where content, often in the form of references, is retrieved and incorporated after the final response has been fully generated, to provide some indication of knowledge to back it. We are particularly interested in this second category because it provides a clearer path to retrieval influencing the generation in productive ways. In this model, improving the retrieval function can be one of the primary means of improving the system's value, which aligns with the goal of creating an effective information retrieval partner for research. This reflects our central thesis: we believe in de-emphasizing the final generated text and instead using generation as a tool to accompany the user's critical and reflective engagement with the retrieved content. In the review by Li et al., the eight works of particular interest to us we had mapped beforehand fall within their citation within generation subcategory of RRG, which specifically maps out works that significantly address this citation-like behavior as a form of interaction. It is also worth noting that Generative Retrieval faces more significant technical hurdles, such as scalability in the face of massive corpora and the propensity for models to hallucinate non-existent document identifiers. Thus, hereafter in this section, we will be more directly discussing systems of the second kind, Reliable Response Generation, which entails the augmentation of LLMs with retrieval systems more closely in the molds we discussed in the preceding sections. This does not mean that some form of ID generation, by itself, is completely absent from RRG systems, as it in fact can also be an important component of RAG systems, and still presents particular technical challenges relating to consistency and avoiding hallucinations.

This consolidation of the field is also helpful to more clearly establish the differences between *generative information retrieval* (Gen-IR) and other similar applications of LLMs. For instance, while works built around the distinct idea of *Generative Recommendation* (GenRec - as reviewed inDeldjoo et al. (2024)) also use LLMs to produce outputs, their goal is typically to suggest items directly. This practice often incorporates methods from Generative Retrieval, where the LLM is responsible for generating IDs that serve as pointers to a catalogue of possible items, and are more often concerned with pointing towards multiple items at once, and in incorporating aspects such as likelihood ratings and other forms of certainty communication (Li et al., 2025; Wang et al., 2024b). By contrast, Gen-IR is specifically concerned with retrieving and presenting or integrating generally textual information in response to a user's query (Li et al., 2025; Wang et al., 2024b). Within RRG and similar commercial kinds of retrieval systems, a key design choice emerges regarding how to present retrieved information. This exists on a spectrum: some systems incorporate information almost seamlessly into a final text, while others treat the retrieved content as a primary layer of information to which the generation refers. A system in the molds we are describing would use generation not just to cite, but to actively accompany the user's interaction with this primary layer of retrieved content, pointing them towards relevant information. The works most relevant to our paper are those that use external mechanisms to explicitly improve the retrieval stage, thereby enhancing the entire system's utility. For instance, some frameworks use fine-grained, sentence-level rewards to train models for better attribution, while others like the AGREE framework fine-tune an LLM to iteratively identify its own unsupported statements and retrieve new information to substantiate them (Huang et al., 2024; Ye et al., 2024). The LLatrieval framework similarly verifies if retrieved documents are sufficient and generates missing-info queries to supplement the knowledge base (Li et al., 2024b). The technical goal of these approaches is to improve the signal-to-noise ratio of the context provided to the LLM, ensuring the generated output is more accurate and useful. These are important considerations that informed us not only on the employment of a particular RAG format, as detailed in Section 5, but also on the exploration of how to incorporate additional data that can improve the results for a particular user.

Some more structured systems incorporate these retrieval optimizations as part of intermediary generation steps, such as by using chains of prompts, or chain of thought, which allows subsequent generations to benefit from this dynamic evaluation of source material. The *Search-in-the-Chain* (*SearChain*) model, for example, has the LLM first generate a complete reasoning plan, which a dedicated information retrieval module then uses to verify claims at each step (Xu et al., 2024). Similarly, the *Hierarchical Graph of Thoughts* (HGOT) framework expands a single query into a multi-layered graph of interconnected sub-queries (Fang et al., 2024). The core idea behind these approaches is the use of sequential, chained queries where the outcome of one step informs the context for the next. This principle of enabling a continuous, context-aware inquiry through multiple evolving queries is a central theme that will be explored further in the design of our own system. Of particular significance are the systems that attempt fuller personalization by directly incorporating data from the user's evolving context and work. Recent studies evaluating *ChatGPT*'s ability to perform personalized retrieval tasks found that while it struggles with high-accuracy ranking compared to specialized systems, it excels at generating coherent, human-judged explanations for its recommendations (Liu et al., 2023). Taking a more structured approach, the *LaMP* benchmark provides a comprehensive testbed for personalization tasks and, critically, proposes a two-step retrieve-then-prompt process to overcome context window limitations (Salemi et al., 2024). In this process, a search component first finds the most relevant examples from a user's vast history, and this small, targeted selection is then used to personalize the prompt fed to the language model.

Moving beyond simply retrieving relevant historical examples, a more sophisticated approach to personalization involves actively structuring this data to give it meaning in the context of the user's ongoing goals. This act of structuring past user data brings it closer to knowledge representation, aligning with our goal of creating an epistemic tool that aids research as a reflective information-seeking activity. For example, the *Cognitive Personalized Search* (*CoPS*) model processes a user's entire search history to infer their personalized query intent, creating a structured profile of their interests and background that is combined with their current activity to describe what they are trying to achieve (Zhou et al., 2024b). Similarly, the *User Interest Journeys* framework uses an LLM to analyze and apply human-readable names to semantic clusters of a user's past interactions, providing a structured summary of their persistent interests (Christakopoulou et al., 2023). The principle of structuring user-provided data is fundamental to our own system, as we will later detail. Other approaches achieve personalization through sophisticated prompt engineering that structures the interaction itself, rather than pre-structuring the historical data. The *Tailored Visions* framework, for instance, rewrites a user's current prompt to incorporate stylistic details retrieved from their most relevant past prompts (Chen et al., 2024). The *BookGPT* framework assembles distinct components into its prompts, such as an injected identity for the model's role and a task description for the goal, to guide the generation process (Zhiyuli et al., 2023). By carefully defining the rules of the interaction, this method ensures the LLM's output is precisely tailored to the task at hand. And, by modeling the user's past behaviors and evolving interests, these memory-based systems move beyond single-shot, transactional queries. As we will show, our proposed design also leverages this technique by dynamically assembling prompts that incorporate not only the user's query but also the structured context derived from their explicit interactions with their research materials.

In table 3.1, we laid out a table of the works cited in this section, reflecting the logic we presented from the broad view of RRG works, to the structured use of user context matching our proposal. The additional dimension of RRG works that leverage multi-step generation is marked in grey.

An important category of related works to this one are those that employ these LLM related practices for academic work specifically. The review by Li et al. mentions a few retrieval focused works in the academic domain, as part of a broader discussion on Domain-specific Personalization, for instance, the *RevGAN*(Li and Zou, 2019) system for generating controllable and personalized user reviews , and works focused on writing assistance like *Pearl* (Mysore et al., 2024), a personalized LLM writing assistant trained on a user's historical documents (Li et al., 2025). However, we present a broader review of academic work related publications in the following section, including a few retrieval focused works that were not yet published when their review was written.

Table 3.1: A table of the works cited in Section 3.1, from the broad view of RRG works to the structured use of user context. Our proposal is marked in underline and bold text. The additional dimension of RRG works that leverage multi-step generation is marked in grey.

RRG works focused on retrieval quality/relevance	
Ye et al., 2024	AGREE framework
Li et al., 2024	LLatrieval framework
User context personalization	
*Liu et al., 2023**	Is ChatGPT a Good Recommender? A Preliminary Study.
*Salemi et al., 2024**	LaMP benchmark
Chen et al., 2024	Tailored Visions framework
Zhiyuli et al., 2023	BookGPT framework
Structured user context personalization	
Zhou et al., 2024	Cognitive Personalized Search (CoPS) model
Christakopoulou et al., 2023	User Interest Journeys framework
Generative retrieval probing	**Our proposal**
Xu et al., 2024	Search-in-the-Chain (SearChain)
Fang et al., 2024	Hierarchical Graph of Thoughts (HGOT)

▨	Works incorporating multi-stage retrieval
*	Overall method benchmark works

3.2 Works on LLMs for academic information seeking

Having established the technical landscape, this section positions our contribution within the specific application domain of academic information seeking. A multitude of works on various forms of information seeking using LLMs in an academic context are focused on aiding in academic literature reviews, similarly to commercial products like *Elicit* and *SciSpace*, with proposals like *PRISMA-DFLLM* (Susnjak, 2023) and *LitLLM* (Agarwal et al., 2024) gaining citation momentum. A literature review of proposals of this kind by Scherbakov et al. Scherbakov et al. (2025) mapped 172 such papers. The majority of those were in the categories of searching for publications, data extraction, and title and abstract screening, with the overarching goal among works mapped by them being in completing parts of a literature review more quickly or easily. The important distinction we make from such tools is our focus on aiding the underlying knowledge work itself, rather than on the completion of a specific, narrow task in the writing of academic papers. Our review will therefore concentrate on a select group of works designed to support the more conceptual and analytical aspects of research. We will begin by examining non-retrieval-heavy analytical tools, followed by the more centrally relevant retrieval-focused systems. To conclude, and to clarify the scope of our proposal as a mixed-initiative system, we distinguish it from more ambitious works that aim for the full automation of the science process.

A more restricted set of works, including a few present in the aforementioned review, seeks to perform information seeking among academic papers solely to provide insights to the researcher. These systems often focus on intermediate, non-final tasks by supporting key parts of the knowledge process. In the literature we reviewed, the most common forms of this support were idea generation, hypothesis formulation, and other forms of analysis such as formulating research questions or identifying research gaps through topic modeling.

We can first examine some exemplar works that provide this support through mechanisms that do not rely heavily on retrieval. As a form of analysis, *CoQuest* (Liu et al., 2024) is a proposed tool designed to assist researchers in formulating research questions by employing a structured conversational agent. It guides users through a process that involves either a breadth-first exploration for a wide range of question variations, or a depth-first exploration to delve into more specific formulations. Another analytical approach is seen in the work of Abd-alrazaq et al. (2024), who employ BERT to perform topic modeling on a corpus of research papers to help discover potential research gaps. In the area of idea and hypothesis generation, *ChatCite*, (Li et al., 2024c), mimics human workflow to produce a comparative summary from a pre-defined collection of papers, while *HypoGeniC* (Zhou et al., 2024a), is a data-driven framework that iteratively generates and refines hypotheses based on their predictive performance on a set of labeled examples.

It is worth noting that any Chat Q&A or simpler Paper Summarization systems that incorporate retrieval, which includes both the commercial LLM chat systems we already mentioned, as well as a series of similar products and works directly positioned for academic work, could be said to be aiding in the knowledge processes related to research, and providing insights to the researcher, but we believe these general purpose unstructured tools do not fall in the purview of our goals. There are, however, works that support the knowledge processes of idea generation and hypothesis formulation by explicitly employing retrieval as a central mechanism. As mentioned in the preceding section, many of these were published after the cutoff date of the systematic review by Li et al. and serve as interesting examples of retrieval-focused systems.

Some of these systems focus on extracting conceptual components from papers to generate novel insights. The work of Li and Ouyang (Li and Ouyang, 2024), for example, investigates how LLMs can uncover and explain relationships between papers by extracting relevant features from a paper and its citations, guided by a user's natural language plan. This idea of deconstructing papers into core components is central to other tools as well. Both *Scideator* (Radensky et al., 2025) and *IdeaSynth* (Pu et al., 2024) are built around the concept of facets, which they retrieve from user-provided literature to allow for interactive recombination and idea development. This approach of using features and facets directly informs the rationale for *GRP*'s retrieval methodology, which is designed to operate on the concept of both explicit and implicit knowledge blocks, which we detail in our study section.

Another group of systems uses highly structured, multi-step retrieval processes, as discussed as part of Section 3.1. The *Chain of Ideas* agent (Li et al., 2024a), for instance, constructs a chronological map of a research topic's evolution by retrieving an anchor paper's references and citations. Others, like *Nova* (Hu et al., 2024), employ an iterative planning and search methodology to retrieve external knowledge, while *MOOSE-Chem* (Yang et al., 2025) and *Literature Meets Data* (Liu et al., 2025) formalize retrieval as an explicit first step to gather inspirations or literature to be synthesized. Our own proposed method builds upon this concept of a multi-step retrieval process, but is uniquely guided by specific kinds of data inputs provided by researchers during their workflow (as detailed in Section 5).

A more recent literature review by Eger et al. (Eger et al., 2025), published very close to the date this work was being finished in April 2025, aids in mapping out these works, placing them in categories very similar to our own, such as *Ideation, Hypothesis Development, and Experimentation* and *Literature Search, Summarization, and Comparison*, as well as other areas covering the broader use of LLMs in academic work like *Text-based Content Generation*, *Multimodal Content Generation*, and *Peer Review.* In their overall conclusions for the field, they denote that "One potential avenue is enhanced personalization which can be achieved by adapting search engines to user preferences, providing more tailored recommendations based on research interests and behavioral patterns" (p.8). They further identify potential in "fostering interdisciplinary collaboration through the integration of AI-powered search systems with other digital tools, such as data visualization platforms and research management software, could facilitate more comprehensive and insightful research outcomes" (p.8), which we believe are points closely related to this work.

There are also works in this space that seek to fully automate the science process using LLMs, some of which very heavily incorporate retrieval. Prominent examples of these more comprehensive systems include *The AI Scientist* (Lu et al., 2024), a framework designed to automate the entire research pipeline, from idea generation and experiment execution to the writing of the final paper. Similarly, recent efforts from major research labs like Google have explored multi-agent systems, such as an *AI co-scientist* that uses a generate-debate-evolve framework to iteratively enhance hypotheses (Gottweis et al., 2025). While these approaches are ambitious, our work is distinct in its focus on supporting and collaborating with the human researcher on the specific, reflective tasks of information seeking and knowledge synthesis, rather than automating the scientific workflow itself. We consider that these fall outside of the scope of this work, as our proposal focuses on a mixed-initiative system designed to augment a researcher's cognitive process rather than achieving full automation.

In Fig. 3.1, we have the LLM works mentioned in this section distributed along a spectrum – from tools for working with literature; to idea, hypothesis and experimentation tools; to works proposing full scientific automation using LLMs. We marked our target with GRP along this same line – a system that leverages the generative aspect to propose points purely to be validated by the user, grounded entirely within selected literature.

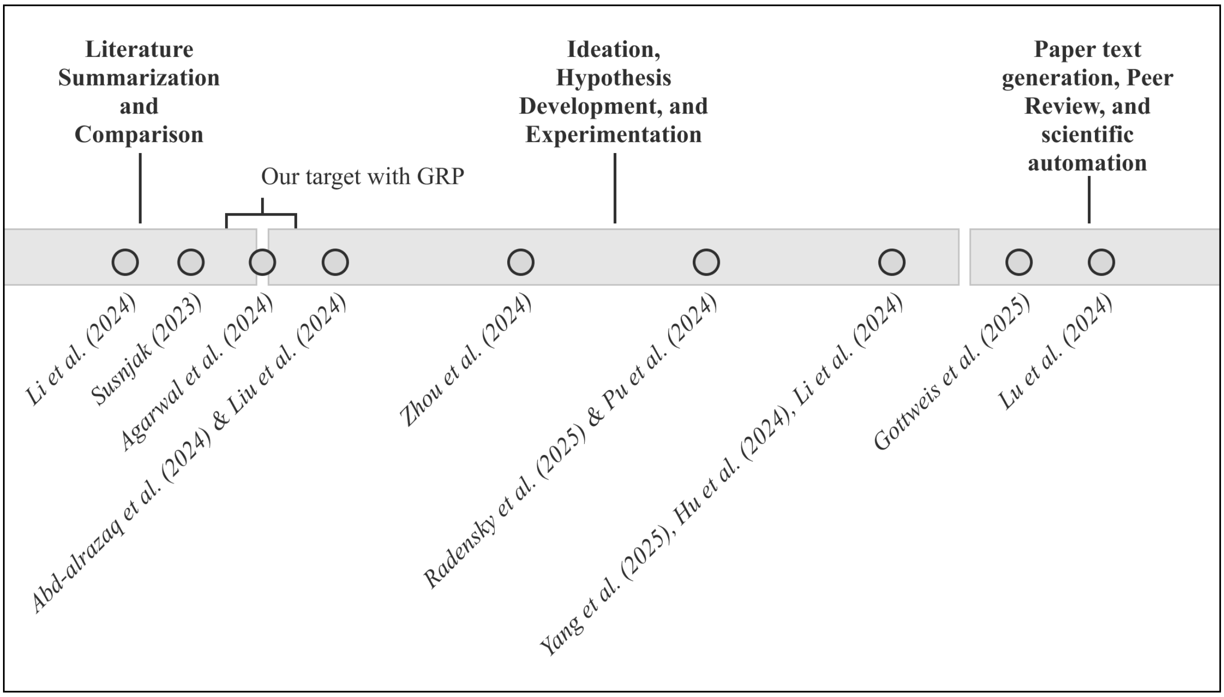

Literature Summarization and Comparison
Our target with GRP
Ideation, Hypothesis Development, and Experimentation
Paper text generation, Peer Review, and scientific automation
Li et al. (2024)
Susnjak (2023)
Agarwal et al. (2024)
Abd-alrazaq et al. (2024) & Liu et al. (2024)
Zhou et al. (2024)
Radensky et al. (2025) & Pu et al. (2024)
Yang et al. (2025), Hu et al. (2024), Li et al. (2024)
Gottweis et al. (2025)
Lu et al. (2024)

Figure 3.1: The works mentioned in Section 3.2, distributed along a spectrum – from tools for working with literature; to idea, hypothesis and experimentation tools; to works proposing full scientific automation using LLMs.

4
Preliminary study

To begin answering our first research sub-question – *What are researchers' current workflows for literature knowledge, and what are the key challenges and opportunities for AI in that environment?* – we conducted a preliminary qualitative study. Central to this investigation was the need to understand how researchers currently perform information seeking and management in their domain, and how their experiences with LLM-based systems and other AI tools, chiefly recommendation algorithms, fit into these workflows. This investigation was designed to ground our work in the authentic practices and perceptions of our target users, providing an empirical foundation for any subsequent design intervention.

4.1
Preliminary study planning

To accomplish this, we performed a qualitative interview study, in which it would be possible to dig more deeply into real researcher's work and perceptions. We chose not to focus specifically on researchers that already make heavy use of LLMs or recommendation tools and instead sought general experiences from researchers in our community.

It is worth noting that not all kinds of researchers work with literature in a set way, and that those in fields of expertise too different from our own might have completely different concerns. In light of this, we limited this study to STEM researchers, from diverse disciplines within. Our selection criteria included individuals spanning both early-career and senior researchers, in both academia and industry, encompassing different environments where research is conducted. This approach was designed to allow us to capture a variety of information-seeking behaviors and attitudes toward the integration of new technologies.

As recommendation systems and information management tools are strongly related to creating connections and drawing similarities from a knowledge corpus, we also investigated how researchers experiences map onto their perceptions on connections and similarities when working with papers. It is worth noting in this regard that computational similarity, while not guaranteed to work in the same way as humans would perceive it, is also a core component of most RAG systems.

4.1.1 Goals

Our goal was to understand researchers' workflows for managing literature knowledge by investigating their current practices, their perceptions of conceptual relationships between texts, and how existing AI tools currently fit,or fail to fit, within this environment. To achieve this, we broke the investigation down into four core sub-questions:

1.1 What is the role of recommendation systems and of LLM-based systems in researchers current work?

1.2 How do researchers work with traditional search systems?

1.3 How do researchers organize and manage academic papers?

1.4 How do researchers perceive connections and similarities among academic papers?

4.1.2 Procedure and materials

Participant recruitment was carried out by targeted publicizing of the study within research groups and professional networks accessible to us, spanning both academic and industry settings. We began by distributing a recruitment and screening form, which collected information on participants' research backgrounds, areas of expertise, and general research context.

The screening form was composed of six questions designed to capture a snapshot of the respondent's research profile. These questions ranged from the duration of their involvement in academic research to their highest completed degree and the context in which they conduct their research work, whether in academia, industry, or other settings. Participants were also asked to provide an open-ended description of their main research area. The full content of the form can be found in Appendix A.

An additional point is that respondents were also asked to indicate how often they work independently versus in collaboration with others, using a 5-point Likert scale to quantify these activities. While not considered for screening, we believe this was an important bit of background information to have before talking to participants, as it might color their perceptions of certain research practices.

The initial pool that responded to our outreach was then expanded through snowball sampling, as we asked participants to recommend other researchers they knew of a similar level, specifically in institutions other than their own. This approach allowed us to achieve a final balanced sample of ten participants. Table 4.1 shows the final distribution of participant profiles accross experience and latest academic degree, further separated by purely academic experience ,and both academia and industry; as well as the distribution of these participants along the different fields.

Table 4.1: Final distribution of participant profiles accross experience, latest academic degree, area, and field.

Participant & Area	**Academic Status**	**Exp. (Yrs)**
Industry Participants		
Computer Science		
P1 - Machine Learning (ML)	Completed PhD	16
P9 - Artificial Intelligence (AI)	Completed PhD	22
Chemistry		
P5 - Materials	Completed Master's	12
P6 - Materials	Completed Master's	12.5
Physics		
P2 - Nanotechnology	PhD in progress	13
Academia Participants		
Computer Science		
P4 - Graphs	Master's in progress	4
P3 - Human-Computer Int. (HCI)	PhD in progress	7
Engineering		
P8 - Civil Engineering	PhD in progress	10
P7 - Electronic Engineering	Completed Master's	5
Physics		
P10 - Particles	PhD in progress	4

For the interviews, the primary material developed was the interview script, which can be found in Appendix B. The script was composed of 31 questions and sub-questions derived from our broader sub-questions, designed to be executed in 45 to 60 minutes in a semi-structured format. The questions were distributed among core topics as follows:

- Sq1 - Role of recommendation systems and LLM based systems in current work - 8 questions
- Sq2 - Traditional search systems - 6 questions
- Sq3- Organizing and managing academic papers - 7 questions
- Sq4- Perceptions of connections and similarities among academic papers - 10 questions

To evaluate and refine the study materials, a pilot session was conducted with an additional researcher fitting the screening profile, also recruited from this general pool.

4.2 Analysis

For the analysis, we first transcribed the interviews, followed by a pass to anonymize them by removing any specific mentions of individuals' names or institutions. We then employed an open coding (Corbin and Strauss, 1990) on the transcriptions. We opted not to adhere to a strict coding methodology, instead focusing on selecting passages that shared common themes and messages that stood out as noteworthy and tied back to our research topics and core goals.

As part of the coding, we maintained a rolling codebook. This served as a working document that evolved throughout the process and was used as a guide during the coding phase. We collected 36 codes at the conclusion of this step.

We then moved to affinity mapping of these codes, grouping them based on shared themes and similarities. As part of these process, we also organized the groupings around our core questions, allowing our conclusions to reflect roughly the same order as the original script. In the following discussion of our findings, specific quotes are numbered when relevant, which can be found in Appendix B.1.

4.3 Discussion

4.3.1 The role of recommendation systems and of LLM-based systems in researchers current work (Sq1)

When discussing recommendation tools for academic papers, participants expressed varied interpretations of what constitutes a recommendation and often sought clarification. This wasn't due to a lack of knowledge or experience but seemingly because **the concept of recommendation in a literature-seeking context is ambiguous**. Most participants (P1, P2, P5, P6, P7, P10) defined it as tools that find papers without direct user intervention, excluding rule-based alerts, while some considered any tool that retrieves literature without a direct query as a recommendation system. However, neither type of tool was widely integrated into participants' workflows, with only minor or occasional use reported, and some dissatisfaction among the heavier users.

Regarding tools that provide alerts for new publications, staying up to date with research was a common goal, which we will discuss separately in a further section, but participants found existing tools inconsistent. While theoretically useful, email alerts were hard to configure and often became overwhelming, ceasing to be helpful after some time (P1, P2, P3, P4, P9). This highlights a **gap between the desire to stay updated and the utility of recommendation tools, compounded by challenges in interpreting recommendations**, even suggesting a sort of bottleneck in interpreting recommendations from these tools (P2, P3, P5, P9) [Q1]. Nevertheless, **recommendations from previously found literature were valued as a way to extend the literature search without additional effort** (P2, P10) [Q4, Q5]. This sentiment was shared by participants who occasionally used similar paper tools in libraries and those who used dedicated paper-from-seed-paper tools (P2, P6, P7). Additionally, some participants, particularly those working in domains where diagrams are common in papers, such as for structures or experiment setups, noted that **image search and image similarity based search have relevance to research contexts**. By going through the images they recognized as coming from academic papers first, they were able to quickly judge if a diagram represented what they were interested in seeing, and iteratively go through similar diagrams presented to find the kind of experiment or structure they were looking for (P2, P8) [Q6].

The frequency of use we noted with LLM-based tools was similar, with most participants having experience with basic tools (P1, P2, P3, P4, P7, P8, P10), and some reservations, and a few participants having attempted to use more complex tools with limited success (P2, P4, P10). However, in the case of LLMs, nearly all descriptions of use were more based on curious testing and isolated use attempts, rather than more pervasive contact. For example, we had no cases of participants that made regular use of bespoke LLM literature search tools (ex: Elicit) in the study, but many participants reported being aware of them. In terms of literature seeking specifically, **many of the participants had attempted to use basic LLM tools for finding papers, with trustworthiness issues**. This mostly came in the form of asking widespread chat tools, such as ChatGPT or Gemini, for literature recommendations, which resulted in the generation of fake papers (meaning plausible but non-existant) in the majority of reported experiences (P3, P10) [Q7, Q8]. This led to discussions of **general untrustworthiness of LLMs as a core concern**. **Participants talked about it as much from principles as from experiences, even equally between a computer science background and other areas**, with conversations surrounding how models work, and the kind of response they were able to give as a result. The general perception among participants was that LLMs were not question answering systems, but text generation systems that were sometimes able to answer questions (P1, P2, P3, P7, P8) [Q9]. While the use of LLMs for direct literature seeking is limited by their untrustworthiness, we observed a clear pattern of **mistrust in generative AI combined with cautious optimism for ideation**. Participants consistently saw utility in the outputs as a general guide or idea, independent of their correctness. For example, even when a model generated nonexistent literature, one researcher found the "made up" titles useful for finding real papers via search tools (P7) [Q11]. This suggests that LLMs can be helpful for providing directions and ideas, even if they cannot reliably point to existing citations (P9) [Q10]. One participant described this effect as "talking to an intern," where the conversation is productive even if the "intern" lacks complete knowledge [Q32].

It is worth noting that **experiences still varied between participants, which might be attributable to the newness of tools.** In the lone case of a participant that consistently used LLM-chat tools every day, he echoed these same trustworthiness concerns, but reported the tools as being very useful to him in data processing, such as reformatting lists of citations. Another (P1) mentioned that LLMs were very helpful for them in completing small tasks, such as creating simple code scripts.

4.3.2 Researcher's information seeking in the literature and use of traditional search systems (Sq2)

Literature seeking in search engines **Nearly all participants used a base tool for literature seeking for its ease of use and a second tool for building detailed queries.** Google Scholar was the preferred general seeking tool in nearly all cases due to its simplicity (P1, P2, P3, P4, P5), while Scopus was favored when it came to constructing more refined queries (P1, P2, P3, P4, P5, P6, P10) [Q13]. Participants commonly (P1, P2, P4, P10) began with Google Scholar to get an overview of the literature but switched to Scopus when they needed more advanced filtering, such as narrowing results by country or applying specific restrictions. This transition was usually prompted when participants had a clearer idea of what they needed after an initial, broader search. Participants talked about starting with broader queries in Google Scholar to understand the field and then used more advanced tools like Scopus to focus their search. This dual-tool strategy is indicative of a broader workflow pattern. While discussing their use of these tools, participants often described their work as part of an **iterative cycle where the acts of seeking, managing, and synthesizing information fuel one another** (P2, P6, P10) [Q12]. The switch from a general tool to a more advanced one for detailed query building represents a key part of the seeking phase in this cycle. Participants noted that these detailed tools were better able to support this iteration, and it is possible that the greater **interpretability they afford has a role to play** (P2, P6). As one participant put it, "Understanding what the queries were... helps me understand more of my field" [Q31]. Scopus, for instance, was mentioned as useful because its controlled query structure made it easier to understand how adjustments changed results, whereas the more keyword-based Google Scholar provided less insight. This iterative process of revising searches, facilitated by the more detailed tool, helped participants narrow down their results in a more structured way.

Regarding how participants approached literature throughout the process of seeking, **established research groups emerged as a robust type of findable, followable, and interpretable information store** (P2, P3). Participants frequently described identifying papers not only by their content but also by recognizing patterns in authorship or institutional affiliation. More experienced researchers (P1, P2, P9) talked more frequently about actively seeking out papers from specific research groups, often identifying a core set of authors or institutions known for high-quality work. While this process was more structured for them, it was more organic for others, particularly less experienced researchers that talked about gradually building an understanding of common threads of authorship while researching a theme. Similarly, the familiarity of a researcher with the theme also affected how they talked about seeking papers from research groups. One participant (P2) specifically emphasized that identifying relevant groups was a reliable way to create a structured set of papers when entering an unfamiliar field, serving as a shortcut to building a foundation of knowledge in a new domain [Q14].

Deeper points about general information seeking We also noted that **participants often used tools in a way to purposefully introduce variety to seeking, and also reflected iteratively on these results** (P1, P6, P8). This included varying their queries in unexpected ways or omitting some of the keywords and seeing how results change, but also using certain tools because they felt like they were less direct results of queries, such as repeating a search from Google Scholar in Google Search, to see if the more general purpose engine would lead to different, interesting results. One participant (P1) mentioned he preferred Scopus in general because he felt it gave him more varied results. Expanding on the above point concerning tools that specifically take in a seed paper and find others from it, we noted among the few (P2, P4) participants that used them regularly that these **active tools for finding papers from seed papers are a form of direct query that is more similar to recommendation**, which might be the origin for the capability of extending the process of seeking we previously mentioned (P2). Participants that made heavier use of tools like this mentioned the usefulness of being able to seek papers through other methods, and then check if the tool was able to, in a sense, follow the same line of reasoning as them. While not based on recommendation algorithms, tools that extract or visualize a paper's citations to help a researcher in navigating them are also a form of seed-paper tools. And while **direct citation and reference analysis is one of the main forms of finding papers from papers, dedicated tools for this purpose were not widely mentioned**, with most participants doing this process manually, or through citations pages in the libraries and seeking tools they already used (P2, P3, P4, P7, P8). However, a common thread among participants when discussing both of these topics is that **judging new literature is easier if you know where to put it, and the why of a citation is relevant**, meaning they mentioned looking for literature related literature from authors they already knew among citations when they believed it would lead to good results, and using the citation as a lead on to why a paper would be interesting to them (P2, P8) [Q15, Q16]. One (P2) of the two participants that did use a dedicated citation analysis tool regularly mentioned that not being able to see as easily why a citation was made from the resulting graph was a disadvantage of the tool. An interesting higher-order ramification of the use of research groups as a form of literature seeking, is that **similar advantages appear to extend to researchers working within established groups, who have access to previously used or produced work as an information store** (P1, P2, P3, P5, P6). Participants with experience inside institutional research groups noted that it provides access to a body of internal work, effectively pre-organized (even if not necessarily structured) and ready for use. This internal body of work - papers, datasets, or other resources - served as a direct information source, often reducing the need to conduct a broader search to start, as they had access to relevant materials through their institutional affiliations. On the point of structure, the relationship between these internal information stores and formalized knowledge bases is worth considering. Even if not structured as formal knowledge bases, the ease of access, and even further the ease of finding connections between the materials through access to the researchers responsible for them suggest that these information stores carry some indication of how they might be used, making them more than just information. These informal knowledge stores could, in some cases, approach the function of a knowledge base, even if they are not necessarily recognized as such. In both internal and external sources, we believe the key feature of interest here is the interpretability and the ease of tracing and following the relevant papers, whether in the form of external research groups publications or internal repositories.

4.3.3 Organizing and managing of academic papers (Sq3)

When discussing the organization of research papers, a key finding was that **organization is personal, but often structured around themes**. All participants (P1-P10) organized their papers using theme sets, whether through tags in reference managers or simple operating system folders. What we mean by this is that papers were grouped according to what researchers felt they were about, which in most cases was the field of publication (P1, P3, P4, P8, P10). This played a defining role in how they later retrieved information, with participants then navigating by these themes or recalling previous projects. Interestingly, most participants used folders of PDF files for their primary organization, using reference managers situationally for tasks like generating bibliographies. A minority (P1, P2, P3) of participants consistently used reference managers to maintain a structured base, enhanced by metadata, but this was not the norm. **The second most reported form of organization is by project (P5, P6, P7, P8), but this presented strong overlap with theme**. Participants frequently placed papers into project-specific folders, but in these reports they tended to be closely aligned with the thematic categories of the projects themselves, suggesting that project organization was largely a part or variation of thematic organization for many of the participants. In some cases, **papers were deferred from immediate organization** (P2), with participants relying on provenance features such as browser history or recent downloads to ensure they could locate them later. This allowed participants to delay processing of found literature, without fear of losing track of useful resources. While this may happen for a multitude of reasons, participants who talked about this discussed papers not fitting in to what they were doing at that moment, but seeming potentially useful later.

When discussing note-taking practices related to working with papers, a common approach among participants was **taking notes about the reason for using a paper** (P9, P10) [Q17, Q18]. Participants often made this note after reading, recording why the paper was relevant to their work. There was a wide spectrum of detail and formality in doing this among participants, with some making these notes as part of the organization of their readings only, while a few participants did this in a structured, consistent way. For example, one participant (P10) always wrote on the PDF file itself how he could use noteworthy elements a paper in his work, such as by adopting the same methodology or data set, or replicating elements of a study's design. Another participant explained that she would document the relevance of the article to her research after reading it, and mentioned discarding papers that did not meet this relevance criteria. This process also extended to shared research projects, where she noted why the paper would be useful for a group project or a collective repository of references as it was being constructed. Similarly, **task-based notes around papers were a common form of personal organization** (P2, P4, P5, P7, P8, P9). These notes often included reminders about actions that needed to be taken in relation to the paper, such as re-reading it or finding additional papers on a similar topic. These varied widely among participants, as did the ways they were recorded. One participant, for instance, preferred to handwrite these reminders in a notebook, without necessarily citing the specific paper within the task note. Despite the diversity of approaches, the underlying goal of maintaining a to-do list connected to their research activities through specific action items concerning literature was the same. Another interesting finding among participants that were heavier note-takers was the use of **concept extraction as a form of integrating knowledge from papers into personal notes** (P1, P2, P3) [Q19, Q20]. The idea of a concept was often fuzzy in these discussions, but generally referred to what the paper had to say about a topic of interest or specific theme. Those who used this approach saw it as a way to synthesize connections between multiple papers by passively identifying shared concepts across different sources, and applying to them the lens of personal interest or utility through their personal notes. In a longer timeframe this method helped them build a cumulative understanding of a particular research area, and one participant mentioned it was also a way for him to share definitions of ideas with others when working in a group. In addition to note-taking, a few participants also reported that **intermediary steps in their own paper writing process writing served as part of their organization**, which they often began early in the reading process (P4, P6) [Q21, Q22]. While not users of dedicated note-taking tools, these participants utilized writing tools such as Overleaf [1] to integrate references into draft or outline sections of their writing, even as they were still exploring relevant literature. Similarly, progress presentations to research

[1] https://overleaf.com/

groups or advisors also functioned as organizational artifacts (P6) [Q22]. These presentations, which required participants to summarize their current reading and cite relevant papers, became a means of organizing and recalling important references later in the research process. Finally, **highlighting texts while reading was widely reported (P2, P3, P4, P10), but no specific habits or methodologies were strongly emphasized**. Most participants mentioned using text highlighting as a way to mark important sections or points in a paper, but they did not exhibit strong or consistent patterns in how they did this.

Deeper points on information seeking in organization and management Beyond note-taking, a major component of turning literature into knowledge is the practice of **extracting personally relevant knowledge blocks from the text** (P2, P9) [Q23, Q24]. Our discussions revealed that researchers do this by taking key pieces of information from the context of a paper and transforming them into usable insights, rather than focusing on a comprehensive understanding of the paper in its entirety. Some participants (P2, P8, P9) emphasized that they rarely engage with every aspect of a paper, instead focusing on these blocks of knowledge—which can be specific sections or even isolated sentences—that align with their personal research interests. The work, therefore, revolves around breaking the paper down into these essential blocks, regardless of whether they are immediately obvious within the document's structure. **Sometimes these blocks are a part of the organization of the literature itself, while at other times they represent what the paper says about a particular concept**. For instance, explicit knowledge blocks might include sections like the introduction, figures, or methodology, which are designed to convey essential information about the study's design or scope. A researcher interested in the methodology of a paper as a whole might navigate directly to that section, extracting it as a distinct block of knowledge. Conversely, non-explicit blocks of information require more active engagement from the reader. In these cases, researchers might look for insights scattered across multiple sections, such as descriptions of the use of a material across different domains present throughout the entire text, or look only for mentions of user perceptions embedded within the evaluation section. In such instances, the researcher must piece together information from these different sections to form a coherent understanding of the concept in question. Importantly, **information expected from these blocks is commonly mentioned as a part of reading decisions, usually after some initial quality-based filtering** (P2, P3, P6, P9) [Q24]. Before diving into the paper, some participants (P1, P3, P10) described going through a preliminary stage of evaluation, which included judging the overall quality of the paper. The abstract, title, and sometimes a quick scan of the text help researchers assess whether the article merits deeper exploration. These were all factors considered before then trying to piece together whether a paper contained the desired information. Beyond that, **sometimes, answers predicted for these blocks of information are part of the goal of seeking - what one expects to find in a specific part of a paper**. Participants often mentioned (P3, P6, P9, P10) approaching literature with certain questions in mind, anticipating that particular sections of a paper would hold the answers they were looking for. For instance, a researcher studying a specific methodology might scan through papers using that methodology to quickly gauge the authors' reflections or conclusions about its effectiveness. As above, the decision to read a paper, can be guided by the expectation that it contains a relevant knowledge block, which might require scanning from various different sections of the text, and beyond that the search itself is not only about identifying a useful paper, but also about locating the specific blocks of knowledge that answer the researcher's questions.

4.3.4
How do researchers perceive connections and similarities among academic papers? (Sq4)

When discussing how participants perceived connections and similarities between academic papers, a central finding was that **connections and similarities are fuzzy, subjective, and implicitly managed**. There was noticeable variation in how these concepts were understood, with no universally agreed-upon definition. These terms were often interpreted based on individual experiences, with participants frequently constructing their own working definitions as we talked. This fluidity underscores the subjective nature of both concepts in the context of academic work, where personal perspectives play a significant role in shaping how researchers relate different pieces of literature. For most participants, **connection between papers consists of references or citations, even if they are not there, but should be**. Many participants viewed the presence of references and citations as the most direct form of connection between papers (P1) [Q25]. However, a notable observation among a few of them was that a connection in this sense could exist even if one paper did not explicitly cite another but, in their view, ought to have done so. This idea of a missing citation still representing a connection was a recurring theme and highlights how connections, even if defined in these terms, are conceptual rather than strictly formalized about the references themselves.

In addition to references and citations, **having the same topic or same authors were also mentioned, frequently by the same participants, as another kind of connection.** Participants often recognized thematic or authorial links between papers as another important form of connection (P3, P9, P10) [Q26]. This suggests that beyond explicit citations, participants also relied on broader patterns such as overlapping research areas or familiar authors to draw connections between different works, tying back to the points concerning authorship groups we noted above. Despite our earlier points about note-taking, **basically no participants stored these connections explicitly, though similar effects were reached due to other organizational practices.** Although participants did not tend to document or track connections between papers deliberately, such as by making notes or keeping records about how papers are related, similar results were often achieved through their organizational habits (P1, P4, P10) [Q29]. For instance, many participants grouped papers into thematic folders or collections, which naturally led to related papers being stored together. In this way, connections were implicitly preserved through the broader management of literature, even if they were not explicitly noted. Perhaps the most notable examples of this are the concept notes we discussed above, which helped keep track of connections indirectly by pooling definitions of a same concept among papers.

When it comes to the idea of similarity, for most participants (but not all), **similarity meant doing things the same way, in the sense of using the same method or practices** (P1, P4, P10) [Q27]. Participants frequently defined similar papers as those employing comparable research methods, design, or practices. In this definition, the focus was on what was in the actual text rather than what the paper, as an object, represented. However in opposition to this, **similarity as different practices but the same theme appeared as well**, where participants also considered papers to be similar if they addressed the same topic or theme and the methods used were different (P2, P4) [Q27]. We note this also suggests an overlap between the ideas of similarities and connections, which seems natural given the nature of these ideas as discussed above. Discussing overlaps of these concepts also opens the door to thinking about how they might be combined, or might be parts of a whole. One participant (P2) gave a notable definition, which in a way combined the ideas of having similar practices and same theme - that similarity between papers entails the possibility of transferring knowledge between them, in whatever way that might be. We can draw a parallel from the differences between these definition's to the difference between metadata and data. Some (P3, P8, P9) of the notable discussions about connections circled around what can be said about a paper, such as who the author was, what the theme was, what citations were included. In turn, some of the most notable discussions about similarity (P1, P6, P10) were around what can be said from a paper, such as what the text said about methods, how it was structured, what kind of knowledge was in it. These two concepts may also overlap, as there are kinds of metadata that can be both included in a file and gleamed from the text itself, such as the theme. But, in direct relation to these concepts as being fuzzy, it is notable that **for participants, both the idea of connection and similarity depend on the researcher themselves** (P6, P7, P8) [Q28]. The notion that these concepts were highly individualized was a common sentiment, and participants emphasized that beyond their own definitions, whether two papers were connected or similar depended on the researcher's perspective. These are working concepts largely guided by personal judgment and experience, rather than adhering to fixed, external criteria.

4.3.5 Other relevant themes and perceptions

When discussing how researchers organize their workflows, **we noticed a soft relation between amount of organization and domain knowledge** (P4, P5, P9). Participants talking about experiences in having little knowledge of a new domain tended to talk about managing a higher volume of papers and adopting stricter self-organization practices as they sought to familiarize themselves with new topics. In contrast, the more experienced researchers, who had already established a strong foundation of domain knowledge, talked more frequently more relaxed approach to organization. While they were not necessarily less organized, they were less concerned with actively maintaining detailed organizational systems. Of course, this was secondary to the amount of organization the researcher's themselves preferred to have. Another factor that emerged was the role of **patience and discipline in picking or sticking to certain practices**. When discussing this, participants (P4, P7, P9, P10) [Q30] acknowledged that while they might adopt certain organizational strategies initially, they struggled to maintain them over time. They attributed this inconsistency to a lack of patience or discipline, which in turn influenced their choice of tools and strategies. Cases of practices that required sustained effort or discipline were talked about being avoided or abandoned, even when discussing experiences like we noted above, where researchers became less organized as they became more familiar with a domain because they felt less pressure to do so.

During the course of this study, we did not attempt detailed goal mapping when exploring participants' workflows. Although participants' research goals naturally influenced their organization and strategies, we did not delve into every specific step of their workflow. Instead, our focus was on the broader patterns and practices that emerged in their responses. However, in descriptions by participants of goals, **specific moments influenced certain seeking strategies and practices**, which we believe are noteworthy (P1, P2, P5, P8, P10). For instance, **when trying to get to know a publication field or venue**, participants often focused on the structural aspects of the literature—how papers in that field were framed, written, and presented (P1, P2, P4, P5). This was particularly the case when the field was not entirely new to them, but the venue or more specific sub-field required a different set of conventions or framing in their writing. Similarly, **when getting to know an entirely new domain**, participants reported focusing on the volume of literature they encountered and aimed to form clusters of knowledge (P1, P2, P9). This involved grouping papers around specific topics to help them structure their understanding of the new area. In both cases, domain knowledge acted as an axis along which organization and literature-seeking behaviors varied, with participants using different strategies depending on their familiarity with the subject matter. Another prominent goal was **staying updated while not performing active seeking**. Many participants reported that staying updated in their field, even when they were not actively searching for new papers, was a common priority (P1, P2, P5, P9). To achieve this, they relied more frequently on recommendation tools that helped them passively gather relevant literature. These tools allowed researchers to remain informed without the time investment required for active literature searches, taking into account challenges we noted in the dedicated section above.

4.4 Considerations

From these findings, we believe there are a multitude of valuable insights that can point us towards a proposal. Chiefly, as a substantial element of our focus is information management among researchers, there are key insights we can leverage in **how participants organized and managed papers**. We observed how even when largely unstructured, such as by being based only on loose PDF files, participants rationale behind their organization and their interactions with it ultimately tied back to their particular needs and concerns, and was the underlying structure behind many of the other findings we noted above.

Interesting examples of this arose when participants were discussing whether **similarities and connections between papers** were a part of their organizational practices, as no participants noted that they did this explicitly, but most participants pointed back to elements of their organization when explaining how they handle these ideas. This also ties back to what we observed on how participants fundamentally used **theme sets as a form of grouping papers in their organization**, and especially how this related to their execution of projects and the way they thought about knowledge in general. In this direction, it was noteworthy that most participants we talked to do not use reference managers, and their interactions their own organization and the practices they adopted were largely spread out across many different applications.

We believe the most fundamental of our findings were the ones around **knowledge elements inside a paper**, and the practices researchers employed for **"turning" papers into knowledge**. These knowledge elements are what researchers extract, process, and transform into actionable knowledge, and we observed how the way researchers manage and interact with these elements is central to how they build understanding of overall themes and of their projects. As these ideas are strongly connected to the retrieval of relevant information while working with academic literature, they can be foundational in defining how specifically we might employ LLMs.

On the **existing use of AI recommendation systems and LLMs**, the justified hesitancy we observed surrounding the trustworthiness of LLMs suggests that researchers may not be comfortable relying on these tools for definitive answers or direct information. However, we observed value in LLM tools as a tool for suggestion, in many of the cases working from starting information and generating a direction that was productive, if imprecise.

4.5
Design Goals

From our formative study, we synthesized four stated design goals that articulate our primary design intent: to create a tool that communicates a partnership model, shifting its role from an authoritative answer-provider to a reflective epistemic partner for researchers in both exploratory and organizational phases of their work. While these four goals represent the foundational architectural framework for the GRP system, many other observations from our study—such as those regarding specific search strategies or the subjective nature of textual connections—also directly informed more granular design choices. The Generative Retrieval Probing (GRP) system, detailed in the next chapter, is our structured answer to the second research question: *What design principles and interaction format can position an LLM as a reflective tool, shifting its outputs from full answers to outputs that invite further inquiry?*. It represents a direct combination of this stated design intent and our broader study findings, built upon the principles outlined below.

DG1: **Shift from providing answers to prompting reflection.** Our finding that researchers value LLMs for ideation but distrust them for definitive facts led to our first foundational principle: a system's primary output should be designed to provoke new questions and suggest novel connections, positioning the AI as a reflective partner.

DG2: **Integrate into existing workflows and organizational structures.** The finding that researchers prefer lightweight, personal, and often unstructured organizational methods led to our second core principle: a successful tool should leverage researchers' existing organizational practices without expecting a fully structured or metadata-rich knowledge base.

DG3: **Ground AI pointers toward a verifiable, user-curated corpus.** To counter the observed mistrust in LLM reliability, our third structuring principle is that the system's outputs—its reflective pointers—must be transparently and directly traceable to the researcher's own curated and trusted literature collection.

DG4: **Leverage natural user interactions as input context.** Our findings on personal organization and note-taking practices suggested our final broad principle: a system should be flexible, treating the natural, reflective actions researchers already perform—such as creating notes and highlights—not merely as annotations, but as explicit signals of interest that can drive the generation of new, contextually relevant insights.

5
Proposed solution: Generative Retrieval Probing

Based on the findings from our preliminary study, we propose Generative Retrieval Probing (GRP), an interaction format for LLM-based tools designed to support the reflective and exploratory nature of academic knowledge work. This format is realized in the form of the prototype system detailed in this work, but its underlying principles are intended to be more broadly applicable. The GRP format is founded on leveraging a researcher's existing organizational practices as a basis for improving how they interact with their literature through the use of LLMs.

While RAG systems are capable of information extraction, our study suggests that general information retrieval does not by itself necessarily aid in knowledge work. The GRP format aims to bridge this gap by combining the extraction capabilities of RAG systems with data from user interactions with their research materials, such as highlights and notes, to contextualize the retrieved information. A tool built on this format can better aid researchers in working with the knowledge blocks (as defined in Section 4.3.3) inside papers, supporting the process of turning papers into knowledge that we observed in our first study. To this end, GRP combines LLM-extracted information with the outputs of user interactions to provide contextually retrieved information, along with LLM-generated suggestions on why that information might be relevant.

While our preliminary study noted that not all participants consistently use reference managers, these applications provide a structured environment for an external system to interact with a researcher's existing organization and their interaction data. They serve as a centralized framework and provide the closest available proxy to a researcher's knowledge base. Thus, our prototype is designed as an LLM tool that connects to a reference manager.

Encompassing these key notions, we propose Generative Retrieval Probing as an interaction format centered on a form of proactive and reactive information retrieval focused on reflection. The core of this format is the user's ability to *probe* their collection through a series of context-aware query and generation processes. This probing process is centered on a proactive-reactive loop. Initially, the system performs a *proactive* analysis of a user's paper collection, using RAG to generate a set of thematic starting points. Subsequently, as the researcher interacts with their literature in natural ways—such as adding new documents, creating highlights, or writing notes—the system offers *reactive* retrieval opportunities. These user actions provide immediate context, allowing the system to generate new, more focused suggestions. This iterative cycle, where the outcomes of one probe inform the next, allows the researcher to continuously refine the focus of the analysis based on their own insights and interests, positioning the system as an imperfect collaborative partner in the exploration process. This aligns with findings from our preliminary study, where participants found value in generative tools that provided imprecise but productive starting points for their own analysis.

For this work, we envision the outcomes of these probes to be AI-generated suggestions, designed specifically to incorporate retrieved content alongside the LLM's generative capabilities. The goal of these suggestions is to function as verifiable hypotheses about thematic connections, prompting the researcher's own reflection and supporting cognitive processes such as synthesis and the discovery of novel connections. The specific anatomy and design of these suggestions are detailed further in Section 5.3.2.

- The system performs an initial, **proactive** analysis of a collection of papers, using retrieval-augmented generation (RAG) on the parsed texts to generate a set of suggestions that hypothesize common themes or elements present across the documents.
- For any given paper or set of papers, a user interacts with the texts as they normally would, by adding new documents, creating highlights, or writing notes. In our prototype, this corresponds to the interaction data synchronized from Zotero.
- As they do this, the system offers **reactive** retrieval opportunities. Using the same RAG principles as the proactive step, the user can trigger the generation of new, focused suggestions based on the immediate context of their recent interactions, with the system leveraging this new user data to build more targeted queries.
- This iterative process, where the outcomes of one probe inform the context for the next, allows the researcher to continuously refine the focus of the analysis based on their own insights and interests. The goal of this interaction is to position the system as an imperfect collaborative partner in the exploration process.

This proposal is strongly rooted in our insights from the existing use of LLMs among study participants. We observed that researchers were not comfortable relying on these tools for definitive answers but in many cases found value in their looser, more generative contributions. This aligns with our view that LLMs should be perceived not as authoritative sources, but as collaborative partners that facilitate exploration. Our proposal for GRP embodies this approach, integrating them in a manner that supports exploratory research without demanding complete trust in their outputs. Fig. 5.1 illustrates on a high level the functioning of the GRP concept.

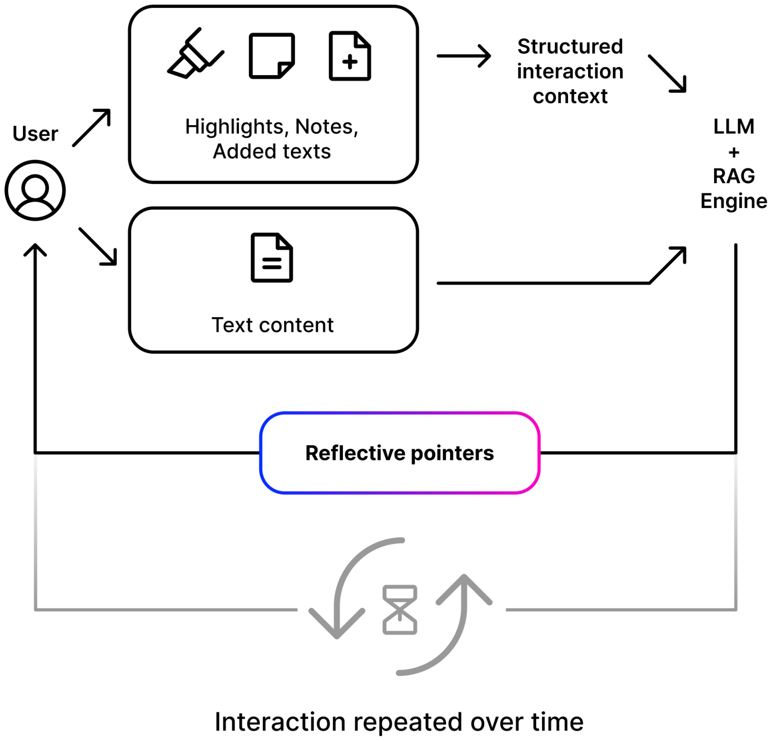

User
Highlights, Notes, Added texts
Text content
Structured interaction context
LLM + RAG Engine
Reflective pointers
Interaction repeated over time

Figure 5.1: The functional architecture of the GRP system's reactive loop. The arrows depict the flow from user engagement with a text to AI-generated 'reflective pointers', which are returned to the user to foster a continuous, iterative interaction.

5.1 Overall description of architecture

The architecture of the Generative Retrieval Probing (GRP) system is composed of two primary components: a backend processing engine designed to run locally on a researcher's machine, and a user interface delivered as a browser extension that operates alongside a reference manager. For this work, we elected to use Zotero [1] as the target reference manager. To facilitate the user

[1] https://www.zotero.org

study, we used the Zotero Web Library UI, which ensured a consistent testing environment and allowed us to create sandboxed collections for participants, regardless of their prior use of the software.

The backend is responsible for the core GRP functionality. It connects to the researcher's Zotero library to synchronize their papers and associated interaction data, such as notes and highlights. These documents are then processed through a self-hosted retrieval-augmented generation (RAG) engine, and a task-queueing system manages the asynchronous operations, from document parsing to the generation of contextual suggestions with a local large language model.

The frontend provides an interactive environment where the researcher can engage with the LLM-generated outputs. A key feature of this interface is the provision of direct, verifiable links from any synthesized insight back to the specific passages in the source documents from which the information was drawn.

The design of this architecture was informed by several technical challenges in the field of generative information retrieval, as Section 3.1. To address the critical issues of hallucination and the need for reliable attribution in RRG systems, our architecture emphasizes verifiability, with the UI cards designed to make the link between a generated claim and its source evidence transparent. To approach the challenge of personalization for complex knowledge work, the backend is architected to integrate not just documents, but also the user's explicit interaction data from their knowledge base. Finally, the decision to use a local, self-hosted engine is a pragmatic response to the challenges of scalability and cost, while also ensuring user privacy for potentially sensitive research. Fig. 5.2 is a detailed diagram of the architecture we described, and which we detail further in the following sections.

5.2
Backend

The backend architecture is fundamentally oriented around the local, self-hosted paradigm. A fully local system means that a user's entire corpus, from their collection of papers to their interaction history, remains entirely within their control. We believe this is a critical consideration for sensitive or pre-publication research. It also provides a pragmatic solution to the challenge of scaling versus cost, important when dealing with LLMs, as it allows the system to process a vast personal library without incurring the financial costs associated with cloud-based services. The limitation with this approach is the lower quality of the models available, as well as the time necessary to process

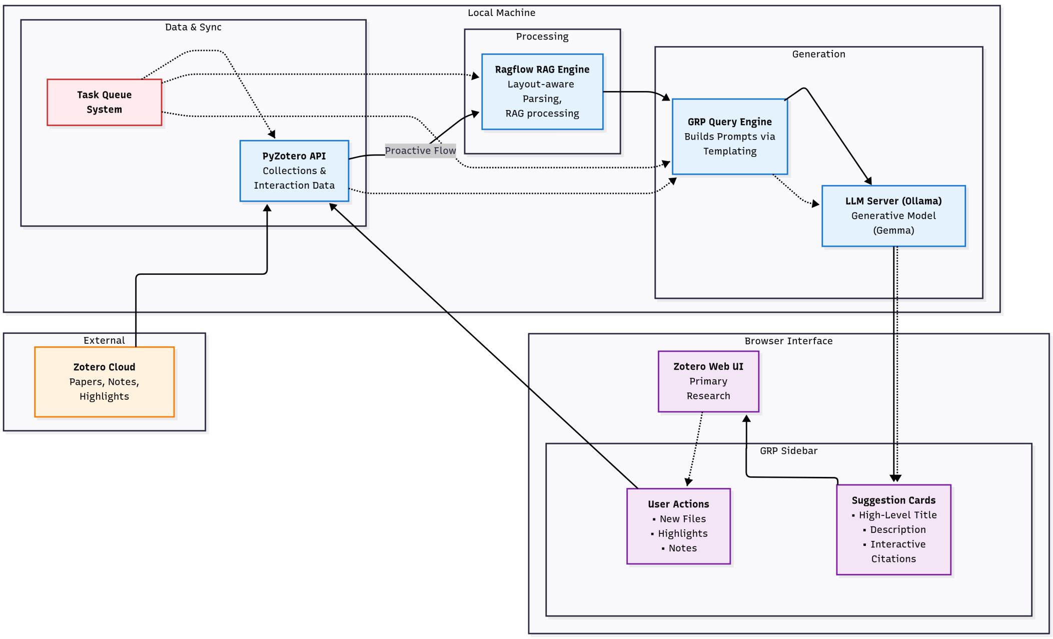


Figure 5.2: The functional architecture of GRP as proposed in this work. In solid arrows, we have the proactive flow that happens independently of user interaction. In dotted arrows, we have the reactive flow from user interaction data to generated results.

the documents and generate responses. For the purposes of this work, we believe that for a research prototype focused on evaluating a novel interaction format, the benefits of user control and privacy outweigh the limitations of using less powerful models.

Given that we were aiming for a system that could be run locally on researchers' machines, it was necessary to set a target set of specifications. For this, we chose a consumer-grade Apple MacBook Pro with an M1 Pro processor and 16GB of RAM as our base model. This hardware configuration is well-suited for running modern quantized language models in the 8-12B parameter range. Quantized models are versions of LLMs where the numerical precision of their parameters has been reduced, significantly lowering their memory footprint with a variable impact on performance. A model in this size range typically occupies between 8 and 12GB of RAM, making it feasible to run on a machine with 16GB of unified memory. While not universally accessible, this hardware profile is a realistic depiction of computers used by researchers, and we used this configuration to perform all initial tests and development of GRP. There is also the possibility that such a system could be self-hosted by a research laboratory for its members, which could be a server or public desktop system equipped with a dedicated GPU with a sufficient amount of VRAM (>=12GB). As noted in the study procedure, the final tests for this work were instead conducted via remote access by the participants to a machine with this second profile, an Intel-based system with an NVIDIA RTX 3060 GPU.

To manage the local execution of LLMs, we first selected Ollama [2] as the model server, which simplifies the process of downloading, managing, and running a variety of open-source LLMs, providing a stable and standardized interface for our application to call upon these models for embedding and generation tasks. When building a self-hosted system, the tools for the application logic beyond the model server exist on a spectrum. This ranges from broad, modular frameworks offering extensive toolkits for custom development, to more focused, end-to-end RAG engines designed for rapid deployment. These engines can function as standalone applications or as specialized components within a larger framework.

To expedite the creation of a user-testable prototype, we also chose to build upon a pre-existing, open-source RAG engine. For this, we chose Ragflow (Fu et al., 2024), an open-source RAG engine, not an LLM itself – but rather the engine that sits between our application and the LLMs to implement the RAG workflow. Ragflow is a complete, end-to-end system specifically dedicated to RAG, offering integrated support for layout-aware document parsing, a built-in re-ranker, and the ability to be self-hosted within our target resource constraints, as well as many other tools dedicated to working with long, structured documents of many kinds, including academic papers. These features made it a competent foundation for our system.

5.2.1 Retrieval-Augmented Generation Architecture

For the GRP interaction format to function, a researcher's library must first be transformed from a static collection of documents into a dynamic, semantically indexed knowledge asset. This transformation is handled by a multi-stage retrieval and generation pipeline, managed by the Ragflow engine, which prepares the text for contextual querying. When a new document is added to a Ragflow Dataset, it is ingested by the engine and undergoes the following steps:

1. **Parsing:** The system first extracts the raw text and structural information from the document. Ragflow uses the deepdoctection (Meyer and Zimmermann, 2022) library for this, enabling it to perform complex layout analysis. This function is important for academic papers as it identifies and preserves structural elements like titles, paragraphs, tables, and figure captions, which helps to maintain the integrity of the various knowledge blocks that researchers value.

[2] https://ollama.com

2. **Chunking:** The parsed text is then segmented into smaller, semantically coherent chunks. Ragflow offers multiple layout-aware chunking methods, including a dedicated mode for academic papers. This approach better respects the document's visual and logical boundaries, preventing sentences from being unnaturally split across columns or paragraphs. The resulting chunks are more likely to represent complete, self-contained ideas, such as a methodological description or a conclusion from a specific section, which aligns with the goal of working with the knowledge blocks identified in our preliminary study.

3. **Embedding:** Each chunk is converted into a numerical vector representation by the embedding model. These vectors capture the semantic meaning of the text, allowing the system to identify thematically related concepts rather than relying on simple keyword matches, a capability that enables the generation of the nuanced connections participants found valuable. These embeddings are stored in a vector database for similarity searching; Ragflow uses InfinityDB, a database specifically designed for efficient, high-dimensional vector retrieval, a necessary component for performing this step accurately on local hardware.

4. **Query-Time Retrieval:** When a prompt is sent to the RAG engine, the query is embedded using the same model. The system then performs a similarity search to find the most relevant chunks from the vector database. A re-ranker model then refines this initial set of results, which improves the "signal-to-noise ratio" of the context provided to the LLM and serves as a direct technical response to the challenge of generating relevant and trustworthy suggestions.

Once this ingestion and indexing process is complete, the system is prepared to execute a probe. The final query process begins with a prompt (either proactively generated or reactively triggered by the user), which is first converted into a vector. This vector is used to retrieve the most relevant text chunks from the indexed documents. These chunks are then re-ranked and compiled into a final context that is prepended to the original prompt before being sent to the Large language model for the final generation of a suggestion.

5.2.2 LLM selection

The selection of specific LLMs for the GRP system was guided by two competing requirements: the need for sophisticated language capabilities to support the proposed interaction format, and the practical computational constraints of our local-first architecture. The system required two distinct types of models: an embedding model to convert text into meaningful numerical representations for retrieval, and a generative model to synthesize retrieved information into coherent suggestions.

For converting text chunks into vector representations, we selected BGE-M3 (BAAI, 2024). This model was chosen for its strong performance on the Massive Text Embedding Benchmark (MTEB) (Muennighoff et al., 2022) at the time of its selection, a standard for evaluating a model's ability to capture semantic meaning. High performance in the key MTEB tasks of retrieval and re-ranking is essential for the GRP system, as it directly impacts the ability to accurately identify relevant knowledge blocks from the source texts and ensure the quality of the context provided to the generative model. This, in turn, helps address the trustworthiness challenges identified in our studies.

For the system's generative component, we considered the trade-offs between smaller, local models and larger, API-accessible ones. While local models are not as powerful as the largest commercial models, recent advancements in quantization have made them a viable strategy for a research prototype. Specifically, quantized models in the 8B to 12B parameter range offer a good balance of capability and resource usage, fitting within the constraints of modern consumer hardware discussed previously. Development began with Google's Gemma 2:9B model and later transitioned to the more capable Gemma 3:12B upon its release (Gemma Team, 2024). This model balances the reasoning capabilities required to synthesize varied retrieved passages into the coherent, thematic suggestions central to the GRP format, with a manageable size for local execution.

This research prototype relies on these pre-trained models without task-specific fine-tuning. However, the potential for future personalization by fine-tuning these models on researcher interaction data presents a compelling avenue for future investigation.

5.2.3 Connection to Zotero interaction data

For the Generative Retrieval Probing (GRP) system to function, it must first connect to the researcher's existing knowledge management environment – as discussed, we consider reference managers to be the closest available proxy to a researcher's organized knowledge base. This section details the process of connecting to a user's Zotero library to synchronize their literature collections and, crucially, the interaction data associated with them.

The GRP system interfaces with a researcher's Zotero account via the platform's API, utilizing the PyZotero library. Authentication is handled using the researcher's user ID and a dedicated API key, which grants access to their personal and group libraries. These credentials are stored locally, providing researchers with direct control over the system's access. This was done so that permission can be revoked at any time by deleting the API key within their Zotero account. However, as mentioned before, to simplify for the final test, we opted to use a single Zotero account where we mirrored the researcher's collections, or manually created them for participants that did not use Zotero before.

Once authenticated, the system creates a local directory structure that mirrors the hierarchical organization of the researcher's Zotero collections. This approach provides a familiar and transparent file-based environment. For each collection selected by the user for inclusion in GRP, the system initiates the following synchronization process:

1. **Collection Metadata Retrieval:** The system fetches and stores metadata associated with the Zotero collection.

2. **Item Identification:** It retrieves a complete record of all items within the collection, including their bibliographic metadata and parent-child relationships (e.g., between a reference and its attached files).

3. **Textual Artifact Discovery:** The system identifies all associated files that contain text, such as PDFs and HTML snapshots, which can be processed by the underlying retrieval engine.

4. **File Download and Mapping:** The identified files are downloaded and stored locally. A mapping file is created to maintain the explicit link between each local file and its parent item in Zotero, preserving the organizational context established by the researcher.

5. **Extraction of Interaction Data:** The system extracts and stores all associated tags, notes, and highlights for each item. This process is fundamental, as this data represents the researcher's direct engagement with the literature.

   - **Tags:** A distinction is made between tags created manually by the user and those generated automatically by Zotero. This is significant because manual tags often represented the theme-oriented groupings and deliberate organizational acts identified in our preliminary study.

  - **Notes:** The system captures the full content of notes, differentiating between those attached to a specific item and standalone notes. Item-attached notes are more likely to contain a researcher's direct reflections on a paper, for example with the observed practice of noting the reason for using a source, while standalone notes may pertain to broader project or thematic goals.
  - **Highlights:** Highlights and their associated comments are extracted along with their precise location data within the document. For example, these may represent explicit markers of what the researcher deemed a knowledge block during their reading.

6. **Dataset Creation:** For this version of the prototype, each Zotero collection is constituted as a separate dataset within Ragflow. This is to ensure that retrieval probes are constrained to the relevant context of a specific project or theme, although in the future we would like to explore more complex cross-collection probes generation.

7. **Content Ingestion:** The downloaded text files are uploaded to their corresponding dataset. The system monitors the subsequent processing pipeline within the RAG engine – including parsing, text chunking, and vector embedding – to ensure propper processing from initial download to final availability for retrieval.

The system then periodically synchronizes with the Zotero API to detect and track changes, such as the addition or removal of tags, notes, files, or highlights. When new files are added, they are ingested into the appropriate RAG dataset. Newly created or modified interaction data is updated in the local storage, making it available for subsequent retrieval probes.

Upon completion, the user's collections, along with their associated textual artifacts and interaction data, are mirrored in the GRP system. This integration allows the subsequent LLM-powered retrieval probes to leverage not only the full-text content of the documents but also the rich ecosystem of metadata and interaction data from the researcher's workflow.

5.2.4. GRP Query Engine

With the researcher's Zotero collection and interaction data synchronized, the GRP Query Engine is responsible for operationalizing the probing process by generating context-aware queries for the RAG engine. This is accomplished using a templating system. The core of this system is a base **system prompt** that is composed with specific instructional blocks and variables drawn from the synchronized Zotero data. This approach provides a consistent and repeatable method for converting the user's rich, multi-faceted context into instructions the LLM can reliably interpret.

A practical challenge in this process is ensuring that any information referenced in the instructions is available for retrieval. To address this, the system performs a two-step availability check: it first consults its local record of downloaded files and then cross-references this with the RAG engine's database of successfully parsed documents. This ensures that any artifact mentioned is fully processed and retrievable. To support the different phases of a research workflow, the system uses distinct compositions for specific contexts. For the initial, proactive analysis of a collection, the following system prompt is composed with the core instructions and data block.

This system prompt (which can be viewed in figure 5.3) was iteratively tested and refined to produce the exact kind of verifiable, hypothesis-driven suggestions we were after. The instructions are deliberately designed to guide the LLM toward identifying thematic connections and grounding them in specific evidence, rather than generating definitive or un-cited answers. The specific relationship between these prompt optimizations and the final anatomy of the generated responses is discussed in detail in Section 5.3.2.

The construction of the final query follows a structured procedure for each collection:

1. First, the system identifies all Zotero items that have passed the availability filtering process and formats their bibliographic metadata into the `{item_titles}` variable. A distinction is maintained between these available items and unavailable ones for logging and transparency.

2. Next, it extracts and formats all user interaction data, including highlights, notes, and tags, preparing them for injection into other template variables like `{unused_highlights}` and `{unused_notes}`. The system processes Zotero's hierarchical structure to capture annotations from child items and combines annotation text with any associated comments.

3. The system then identifies any unused content – highlights, notes, tags, and files that have been added since the last probe was generated. An enhanced synchronization system maintains metadata structures that track this new content, which is the core technical mechanism that enables the system's reactivity. For subsequent interactive generations, the titles of previously generated suggestions are also included in the

```
* Analyze the texts to identify sections with significant
    thematic overlap across the entire collection.
* Respond with a list of high-level themes common to the
  collection, along with notable similarities that
  connect some or all of the texts.
* Identify as many distinct themes and similarities as
  possible.
* For each item in your list, ensure you retrieve all
  relevant text chunks from the source documents that
  support it.
* The user's research goal for this collection is not
  stated and should be inferred from the content of the
  provided documents.
* Respond only with your list of suggestions. For each
  suggestion, provide citations linking to the relevant
  supporting passages in the source texts.
* Ensure that the cited passages are both directly
  relevant to the suggestion and varied in their
  evidence.
* Synthesize these findings into a concise set of
  suggestions that captures all relevant points without
  being redundant.
* Do not include any text before or after your list;
  focus exclusively on the suggestion items.

The following are the available texts in the dataset,
   they must all be considered for your response:
{item_titles}
```

Figure 5.3: System Prompt & Core Instructions

`{kept_headers}` variable to provide context. The LLM can then decide whether to update these themes, use them as context, or ignore them.

4. The appropriate prompt composition is then applied. These compositions are built from Markdown templates that use placeholders for the data variables.

5. Finally, the rendered query is sent to the RAG dataset associated with the collection, and a persistent assistant instance within Ragflow is located or created for that dataset to maintain conversational context over time. Both the query and the generated response are stored for the user.

The outcome of this procedure is a query that combines the full breadth of a collection's documents with the researcher's specific, timely interactions. This incremental approach, which focuses only on new or modified user context, allows the system to react to the researcher's workflow without the need to reprocess entire collections.

5.2.4 Task Queueing System

To manage the interconnected and often long-running processes described in the preceding sections, the GRP system is built on a task-queue architecture that serves as the backend's orchestration layer. This design is well-suited for handling the various asynchronous operations, such as synchronizing with Zotero, parsing documents, and generating new suggestions, which can vary significantly in duration. The queue manages these long-running background processes in an orderly, non-blocking manner.

For persistence, the system uses straightforward, human-inspectable JSON files, a practical choice for a research prototype. These files maintain queues for immediate, long-running, and failed tasks, which are processed at regular intervals by a centralized background worker. A key consideration in this design is the management of computational resources for the local models. To this end, resource-intensive operations, such as document parsing by the RAG engine, are processed sequentially to manage memory consumption and prevent system overloads.

The entire GRP workflow is therefore broken down into a series of modular, specialized tasks. This structure makes the process easier to manage and debug, as the successful completion of one step triggers the next. The main task types are:

- **Setup and Synchronization Tasks:** These tasks handle the initial configuration. `User Setup Tasks` initialize a user's local environment, while `Collection Setup Tasks` create the necessary folder structures and corresponding datasets in the RAG engine. `Download Tasks` retrieve files from Zotero, and upon success, create `Upload Tasks` to transfer the files to the RAG engine.
- **RAG Engine Interaction Tasks:** This group manages the core processing of textual artifacts. `Upload Tasks` include logic to detect duplicates and initiate parsing. The subsequent `Parse Tasks` handle the document chunking. Upon completion of all parsing for a collection, an `Assistant Creation Task` establishes the persistent conversational agent for that dataset.
- **Probe Generation and Execution Tasks:** These tasks execute the generative retrieval itself. `Prompt Generation Tasks` use the templating

system to build the contextualized prompts, and `Prompt Send Tasks` then execute these prompts against the RAG engine's chat endpoints, storing the results.

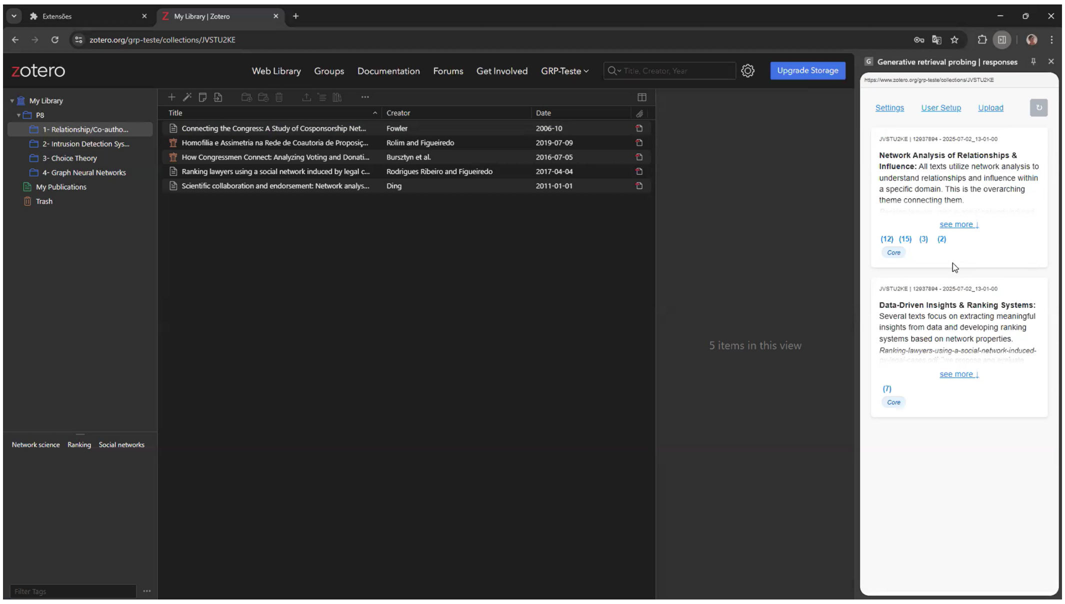
- **System Management Tasks:** To ensure robustness, `Monitor Tasks` track the progress of long-running operations like parsing. Once an operation is complete, they trigger `Cleanup Tasks`, which remove the completed job from the queue and initiate any dependent tasks.

This modular, queue-based architecture provides the necessary structure to manage the multi-stage GRP workflow. It allows the system to operate reliably in the background, translating user interactions and a dynamic collection of literature into the context-aware suggestions that result from the probing process.

5.3 UI Design

The user interface (UI) for the Generative Retrieval Probing system serves as the primary point of interaction between the researcher and the backend's engine. It manifests as a browser extension that functions as a sidebar, designed to augment the researcher's primary reference manager, in this case, the Zotero web UI. By integrating with the reference manager, the GRP system is designed to treat this existing environment as a proxy for the researcher's knowledge base. This approach follows one of our foundational design goals: to integrate into existing workflows and organizational structures (DG2). Consequently, user activities within this space, such as annotating documents or adding new documents, can be interpreted by the backend as signals of evolving research interests; this mechanism operationalizes another of our core design goals, which is to leverage natural user interactions as input (DG4).

The system's main interface is a sidebar that appears alongside the reference manager window (this can be seen in Fig. 5.4). This placement is intended to create a dedicated space for reflective engagement that can support, rather than disrupt, the user's established workflow. At the top of the sidebar, a Settings icon provides access to configuration for debugging, and a Refresh icon allows for manual content updates. This icon notifies the user of newly available suggestions by turning green, providing unobtrusive feedback on the backend's asynchronous processing. The main body of the sidebar presents a scrollable list of cards, each containing a distinct, AI-generated suggestion related to the active Zotero collection.



Title	Creator	Date
Connecting the Congress: A Study of Cosponsorship Net...	Fowler	2006-10
Homofilia e Assimetria na Rede de Coautoria de Proposiç...	Rolim and Figueiredo	2019-07-09
How Congressmen Connect: Analyzing Voting and Donati...	Bursztyn et al.	2016-07-05
Ranking lawyers using a social network induced by legal c...	Rodrigues Ribeiro and Figueiredo	2017-04-04
Scientific collaboration and endorsement: Network analys...	Ding	2011-01-01

JVSTU2KE | 12937894 - 2025-07-02_13-01-00

Network Analysis of Relationships & Influence: All texts utilize network analysis to understand relationships and influence within a specific domain. This is the overarching theme connecting them.

(12) (15) (3) (2)

JVSTU2KE | 12937894 - 2025-07-02_13-01-00

Data-Driven Insights & Ranking Systems: Several texts focus on extracting meaningful insights from data and developing ranking systems based on network properties.

Ranking-lawyers-using-a-social-network-induced-

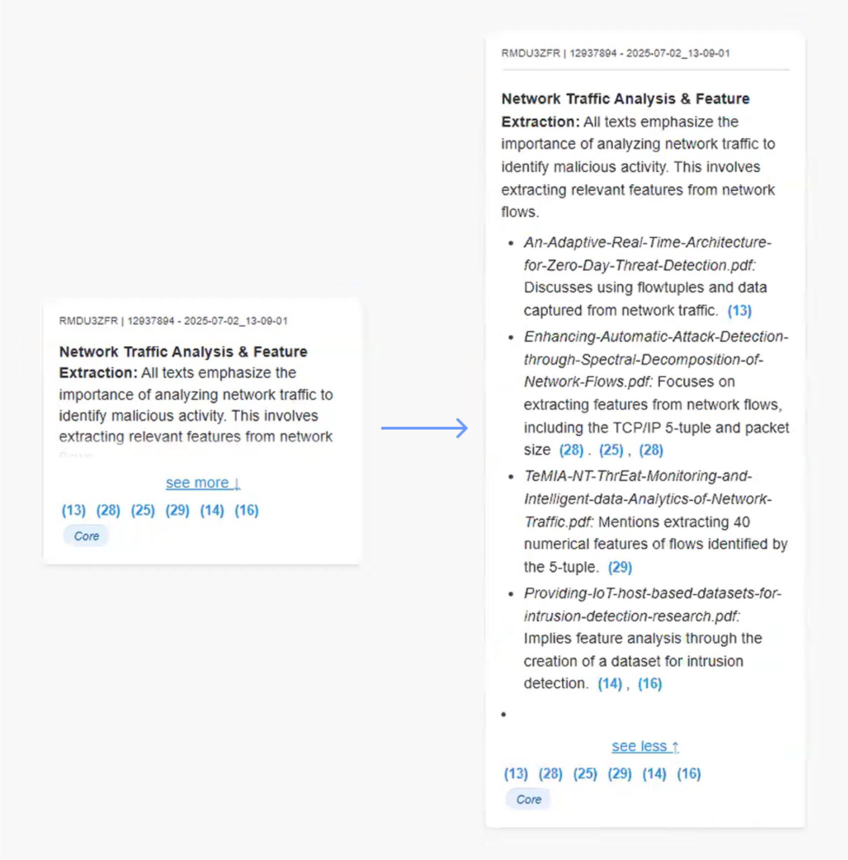
(7)

Figure 5.4: A screenshot of the GRP prototype's user interface. It shows the Zotero web library on the left and the GRP sidebar on the right. The sidebar contains two suggestion cards.

Each card is structured as a self-contained unit of information, representing a hypothesis generated by the GRP system to prompt further inquiry (An example of card in both collapsed and expanded formats can be seen in Fig. 5.3). The card presents its information in a consistent format, beginning with a bolded **High-Level Title** that posits a common thread, followed by a **Short Description** that elaborates on the theme. The main body consists of a **Text-by-Text Breakdown**, which grounds the theme in specific evidence by detailing how it may manifest in the source papers.

A key feature, designed to address the need for verifiability and the issue of LLM untrustworthiness identified in our studies, is the inclusion of interactive, numbered citations within each card. These buttons link to the specific passage, image, or table in the source document that informed the summary. This function is fundamental to the tool's intended role as a reflective aid, allowing the researcher to inspect the basis for the AI's claims and understand the underlying reason for a connection—a practice our participants noted as highly relevant. Clicking a citation opens the source PDF in a modal window that scrolls to the relevant passage, a design intended to support the extraction of knowledge blocks without losing the source context (An example result of this interaction can be seen in Fig. 5.5).

To make the feedback loop between user action and AI generation more transparent, suggestions are categorized with colored tags that signify their origin (as can be seen in Fig. 5.6.)



RMDU3ZFR | 12937894 - 2025-07-02_13-09-01

Network Traffic Analysis & Feature Extraction: All texts emphasize the importance of analyzing network traffic to identify malicious activity. This involves extracting relevant features from network flows.

see more ↓

(13) (28) (25) (29) (14) (16)

Core

RMDU3ZFR | 12937894 - 2025-07-02_13-09-01

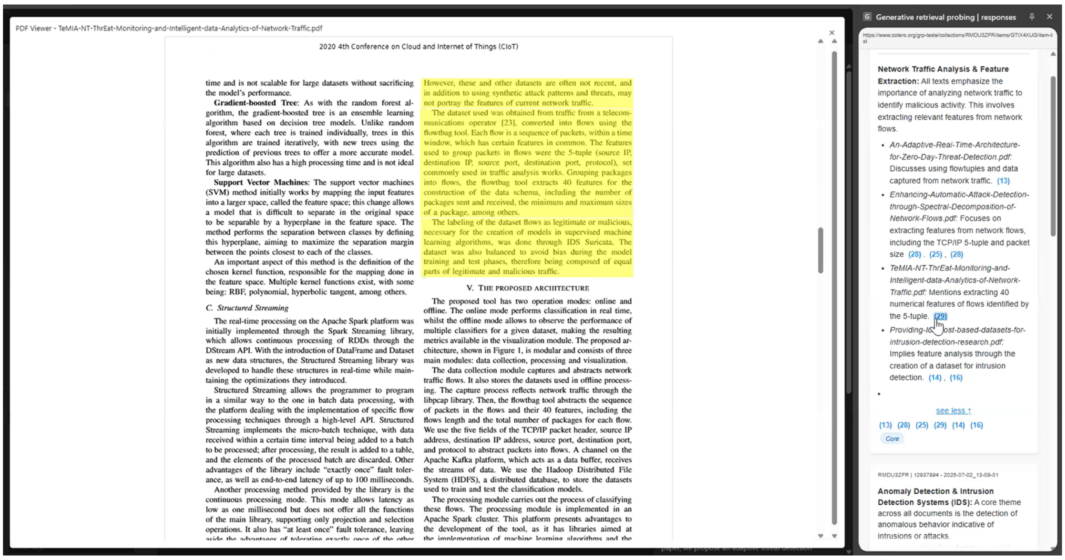
Network Traffic Analysis & Feature Extraction: All texts emphasize the importance of analyzing network traffic to identify malicious activity. This involves extracting relevant features from network flows.

- *An-Adaptive-Real-Time-Architecture-for-Zero-Day-Threat-Detection.pdf:* Discusses using flowtuples and data captured from network traffic. (13)
- *Enhancing-Automatic-Attack-Detection-through-Spectral-Decomposition-of-Network-Flows.pdf:* Focuses on extracting features from network flows, including the TCP/IP 5-tuple and packet size (28) . (25) , (28)
- *TeMIA-NT-ThrEat-Monitoring-and-Intelligent-data-Analytics-of-Network-Traffic.pdf:* Mentions extracting 40 numerical features of flows identified by the 5-tuple. (29)
- *Providing-IoT-host-based-datasets-for-intrusion-detection-research.pdf:* Implies feature analysis through the creation of a dataset for intrusion detection. (14) , (16)
-

see less ↑

(13) (28) (25) (29) (14) (16)

Core



PDF Viewer - TeMIA-NT-ThrEat-Monitoring-and-Intelligent-data-Analytics-of-Network-Traffic.pdf

2020 4th Conference on Cloud and Internet of Things (CIoT)

time and is not scalable for large datasets without sacrificing the model's performance.

Gradient-boosted Tree: As with the random forest algorithm, the gradient-boosted tree is an ensemble learning algorithm based on decision tree models. Unlike random forest, where each tree is trained individually, trees in this algorithm are trained iteratively, with new trees using the prediction of previous trees to offer a more accurate model. This algorithm also has a high processing time and is not ideal for large datasets.

Support Vector Machines: The support vector machines (SVM) method initially works by mapping the input features into a larger space, called the feature space; this change allows a model that is difficult to separate in the original space to be separable by a hyperplane in the feature space. The method performs the separation between classes by defining this hyperplane, aiming to maximize the separation margin between the points closest to each of the classes.

An important aspect of this method is the definition of the chosen kernel function, responsible for the mapping done in the feature space. Multiple kernel functions exist, with some being: RBF, polynomial, hyperbolic tangent, among others.

C. Structured Streaming

The real-time processing on the Apache Spark platform was initially implemented through the Spark Streaming library, which allows continuous processing of RDDs through the DStream API. With the introduction of DataFrame and Dataset as new data structures, the Structured Streaming library was developed to handle these structures in real-time while maintaining the optimizations they introduced.

Structured Streaming allows the programmer to program in a similar way to the one in batch data processing, with the platform dealing with the implementation of specific flow processing techniques through a high-level API. Structured Streaming implements the micro-batch technique, with data received within a certain time interval being added to a batch to be processed; after processing, the result is added to a table, and the elements of the processed batch are discarded. Other advantages of the library include "exactly once" fault tolerance, as well as end-to-end latency of up to 100 milliseconds.

Another processing method provided by the library is the continuous processing mode. This mode allows latency as low as one millisecond but does not offer all the functions of the main library, supporting only projection and selection operations. It also has "at least once" fault tolerance, leaving aside the advantages of tolerating exactly once of the other

However, these and other datasets are often not recent, and in addition to using synthetic attack patterns and threats, may not portray the features of current network traffic.

The dataset used was obtained from traffic from a telecommunications operator [23], converted into flows using the flowtbag tool. Each flow is a sequence of packets, within a time window, which has certain features in common. The features used to group packets in flows were the 5-tuple (source IP, destination IP, source port, destination port, protocol), set commonly used in traffic analysis works. Grouping packages into flows, the flowtbag tool extracts 40 features for the construction of the data schema, including the number of packages sent and received, the minimum and maximum sizes of a package, among others.

The labeling of the dataset flows as legitimate or malicious, necessary for the creation of models in supervised machine learning algorithms, was done through IDS Suricata. The dataset was also balanced to avoid bias during the model training and test phases, therefore being composed of equal parts of legitimate and malicious traffic.

V. The proposed architecture

The proposed tool has two operation modes: online and offline. The online mode performs classification in real time, whilst the offline mode allows to observe the performance of multiple classifiers for a given dataset, making the resulting metrics available in the visualization module. The proposed architecture, shown in Figure 1, is modular and consists of three main modules: data collection, processing and visualization.

The data collection module captures and abstracts network traffic flows. It also stores the datasets used in offline processing. The capture process reflects network traffic through the libpcap library. Then, the flowtbag tool abstracts the sequence of packets in the flows and their 40 features, including the flows length and the total number of packages for each flow. We use the five fields of the TCP/IP packet header, source IP address, destination IP address, source port, destination port, and protocol to abstract packets into flows. A channel on the Apache Kafka platform, which acts as a data buffer, receives the streams of data. We use the Hadoop Distributed File System (HDFS), a distributed database, to store the datasets used to train and test the classification models.

The processing module carries out the process of classifying these flows. The processing module is implemented in an Apache Spark cluster. This platform presents advantages to the development of the tool, as it has libraries aimed at the implementation of machine learning algorithms and the

Generative retrieval probing | responses

https://www.zotero.org/grp-teste/collections/RMDU3ZFR/items/GTIX4XUG/item-list

Network Traffic Analysis & Feature Extraction: All texts emphasize the importance of analyzing network traffic to identify malicious activity. This involves extracting relevant features from network flows.

- *An-Adaptive-Real-Time-Architecture-for-Zero-Day-Threat-Detection.pdf:* Discusses using flowtuples and data captured from network traffic. (13)
- *Enhancing-Automatic-Attack-Detection-through-Spectral-Decomposition-of-Network-Flows.pdf:* Focuses on extracting features from network flows, including the TCP/IP 5-tuple and packet size (28) . (25) , (28)
- *TeMIA-NT-ThrEat-Monitoring-and-Intelligent-data-Analytics-of-Network-Traffic.pdf:* Mentions extracting 40 numerical features of flows identified by the 5-tuple. (29)
- *Providing-IoT-host-based-datasets-for-intrusion-detection-research.pdf:* Implies feature analysis through the creation of a dataset for intrusion detection. (14) , (16)
-

see less ↑

(13) (28) (25) (29) (14) (16)

Core

RMDU3ZFR | 12937894 - 2025-07-02_13-09-01

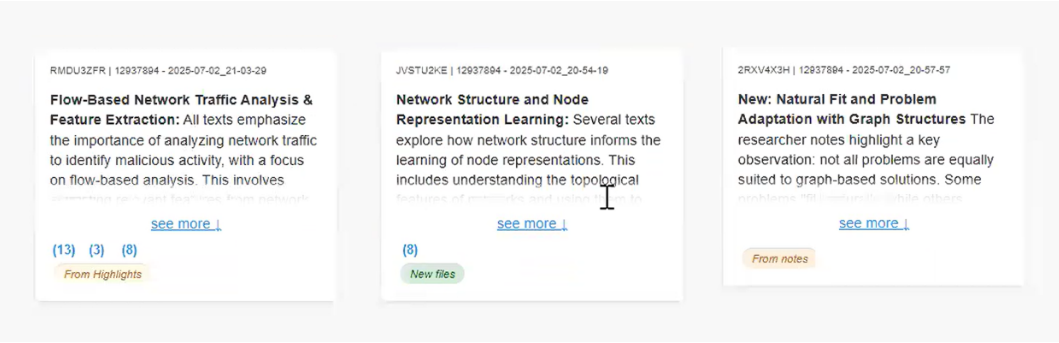
Anomaly Detection & Intrusion Detection Systems (IDS): A core theme across all documents is the detection of anomalous behavior indicative of intrusions or attacks.

Figure 5.5: A screenshot showing the PDF viewer modal. A specific passage in the PDF is highlighted, corresponding to a citation clicked in a suggestion card.

- **Core:** A blue tag indicates suggestions from the tool's initial, proactive analysis of the collection.
- **From New Files:** A green tag marks suggestions generated reactively after a researcher adds a new document.
- **From Highlights:** A yellow tag identifies suggestions derived from the

user's own highlights.

- **From Notes:** An orange tag distinguishes suggestions generated from a researcher's collection-level note, using their written thoughts as a high-level directive.



RMDU3ZFR | 12937894 - 2025-07-02_21-03-29

Flow-Based Network Traffic Analysis & Feature Extraction: All texts emphasize the importance of analyzing network traffic to identify malicious activity, with a focus on flow-based analysis. This involves extracting relevant features from network

JVSTU2KE | 12937894 - 2025-07-02_20-54-19

Network Structure and Node Representation Learning: Several texts explore how network structure informs the learning of node representations. This includes understanding the topological features of networks and using them to

2RXV4X3H | 12937894 - 2025-07-02_20-57-57

New: Natural Fit and Problem Adaptation with Graph Structures The researcher notes highlight a key observation: not all problems are equally suited to graph-based solutions. Some problems "fit naturally" while others

Figure 5.6: An image showing examples of the four types of suggestion cards, each with its corresponding colored tag: Core (blue), From New Files (green), From Highlights (yellow), and From Notes (orange).

This set of features helps operationalize the GRP format by enabling the system to be reactive to user input. When a user provides new context, the tool detects this and presents an opportunity to generate new suggestions via distinctly colored banners that appear in the sidebar.

5.3.1 User Interface Architecture

The GRP user interface is architected as a multi-component system, consisting of a Chrome browser extension that uses the native `sidePanel` API. This panel contains an `iframe` that embeds a standalone Flask [3] web application, allowing it to interface with web-based reference managers like the Zotero Web Library without modifying the host application. The extension uses the chrome.tabs API to detect the active Zotero collection ID from the URL of the active page and passes it to the embedded web application.

The frontend, built with vanilla JavaScript and without a major framework, communicates asynchronously with the locally-hosted Python backend through fetch requests to a series of API endpoints for actions like creating users, initiating background tasks, and retrieving PDFs for display. The UI is then updated using a combination of server-side processing for the response cards and client-side logic, using the marked.js [4] library for Markdown rendering. The frontend periodically polls status endpoints every 30 seconds. This

[3] https://flask.palletsprojects.com/

[4] https://marked.js.org/

live update system is designed to handle failures with continued polling and to preserve user state, such as scroll position, during content updates, however no live updating of the sidebar contents was implemented to avoid disrupting participant's interaction with the cards during the tests. Thus, the content of the sidebar only updated when the refresh button is clicked.

Due to technical limitations of this approach, it was not feasible to highlight the relevant PDF text for each citation, when clicked, directly in the Zotero Web PDF viewer. Thus, a PDF viewer modal is rendered and injected into the main webpage view in a modal, which can be easily close to return to the main Zotero library. This viewer, which enables coordinate-based highlighting, is powered by the PDF.js [5] library.

5.3.2 Design of Generated Responses

The design of the generated responses is central to the function of the Generative Retrieval Probing tool. Informed by our preliminary study, where researchers indicated that generative tools could be useful for providing productive starting points, the responses are framed as verifiable hypotheses rather than as definitive answers. They are intended to act as prompts for reflection, inviting the researcher to validate and interpret the suggested connections. This serves our most central, structuring design goal: to shift the system's role from providing answers to prompting reflection (DG1).

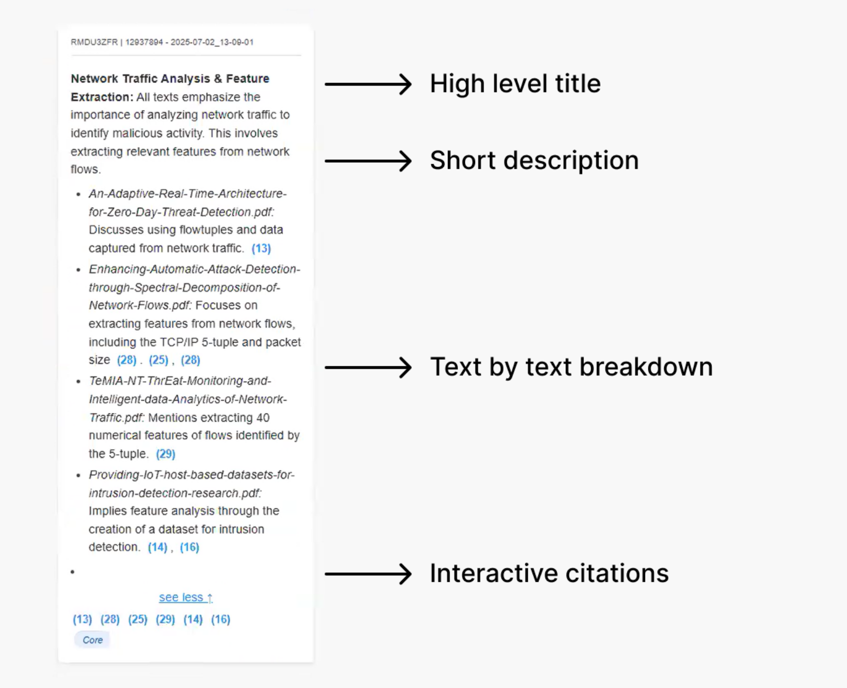
The generation process can result in a hierarchical ordering of suggestions. The system often first generates cards describing broader themes, followed by cards on more specific sub-themes supported by a smaller subset of documents. This can create a progression from a general overview to a more detailed analysis, a common pattern in literature exploration.

The anatomy of the response, embodied in the suggestion card, is structured to bridge abstract concepts and concrete evidence, in an attempt to address the challenges of LLM trustworthiness and verifiability. A detailed breakdown on the anatomy of the cards can be seen in Fig. 5.7.

- High-Level Title: A concise statement summarizing the card's theme. This is designed to act as a prompt for a researcher's own reflective thinking, presenting a potential thematic link that they can then validate or explore further.
- Short Description: A brief paragraph elaborating on the title, providing immediate context for the suggested theme.

[5] https://mozilla.github.io/pdf.js/

- Text-by-Text Breakdown: The main body of the card, which grounds the theme with specific evidence from the source papers. This section aims to address the tension between generality and specificity by providing quotes or summaries that illustrate the theme's presence in cited documents.
- Interactive Citations: Numbered, clickable buttons linking each piece of evidence to its location in the source PDF. This is the core mechanism for verifiability, designed to mitigate issues of LLM inaccuracy and reduce the cognitive burden of verification. This feature is a direct implementation of our third core design goal, which is to ground all AI pointers in a verifiable, user-curated corpus (DG3). It directly supports the researcher's practice of extracting and assessing knowledge blocks by making the system's reasoning more transparent and supporting the user's role in validating the information.



RMDU3ZFR | 12937894 - 2025-07-02_13-09-01

Network Traffic Analysis & Feature Extraction: All texts emphasize the importance of analyzing network traffic to identify malicious activity. This involves extracting relevant features from network flows.

- *An-Adaptive-Real-Time-Architecture-for-Zero-Day-Threat-Detection.pdf:* Discusses using flowtuples and data captured from network traffic. (13)
- *Enhancing-Automatic-Attack-Detection-through-Spectral-Decomposition-of-Network-Flows.pdf:* Focuses on extracting features from network flows, including the TCP/IP 5-tuple and packet size (28) . (25) , (28)
- *TeMIA-NT-ThrEat-Monitoring-and-Intelligent-data-Analytics-of-Network-Traffic.pdf:* Mentions extracting 40 numerical features of flows identified by the 5-tuple. (29)
- *Providing-IoT-host-based-datasets-for-intrusion-detection-research.pdf:* Implies feature analysis through the creation of a dataset for intrusion detection. (14) , (16)
-

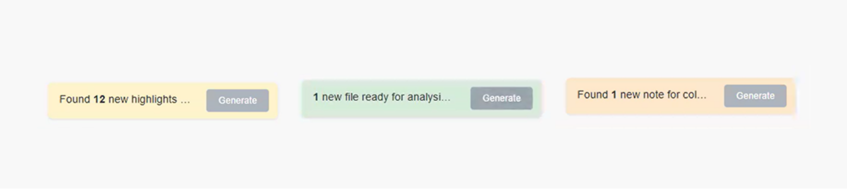
Figure 5.7: An annotated screenshot showcasing the different parts of a suggestion card: a high-level title, a short description, a text-by-text breakdown, and interactive citations.

This highly structured format is a direct result of the specific instructions provided in the system prompt, as detailed in section 5.2.4. Each component of the suggestion card is a fulfillment of a command in the prompt in a repeatable consistent format.

5.3.3 Design of Interactive Generation Features

In line with our fourth broad design goal to leverage natural user interactions as input (DG4), the interactive generation features are the primary mechanism through which the GRP format moves from a passive display of information toward functioning as a more user-directed epistemic tool. These features allow the researcher to probe their collection based on their own actions, creating a reflective loop where user input can inform the AI's analytical focus. Each feature is designed to support a different mode of inquiry, aligning with behaviors observed in our studies.

The design is centered on three user actions: adding new files, creating highlights, and writing notes. After detecting one of these actions, the system presents a reactive banner in the UI (the three types of banners can be seen in Fig. 5.8).



Found **12** new highlights ...

1 new file ready for analysi...

Found **1** new note for col...

Figure 5.8: A screenshot showcasing the different generation banners

This invites the user to generate suggestions based on this new context, operationalized through distinct prompt compositions:

1. **Generating from new files:** This feature is designed to support the incremental nature of research. It directly aligns with the finding from our preliminary study that researchers organize their work into "theme sets" and project-specific folders, gradually building a corpus. This feature allows them to immediately see how a new piece of literature might connect with or alter the thematic structure of their existing collection, supporting an established organizational workflow.

2. **Generating from highlights:** This feature is designed to allow the analysis to emerge directly from the researcher's granular selections within the text. It builds on the knowledge blocks insights from our first study, in connection to the practices we observed in highlighting. This feature is designed to leverage the researcher's own sense-making process by treating their annotations as explicit pointers to concepts they deem important, using them as seeds for discovering related ideas.

3. **Generating from a note:** This feature is designed to enable a hypothesis-driven inquiry guided by the researcher's own high-level thinking. Our preliminary study showed that researchers often approach the literature with an expectation of what they want to find. This feature is designed to formalize that behavior, allowing a user to articulate a research question or theme in a note and use the system to test it against the collection.

These features provide the affordances for the user to guide the RAG engine's focus, with the goal that the generated outputs can become more contextually relevant responses to the researcher's analytical goals. Each feature is powered by a prompt where the base system prompt is composed with additional instructional blocks and data from the user's most recent interaction. The all-encompassing structure for these interactive prompts is displayed in Fig. 5.9.

5.4 System Walkthrough Scenario

To contextualize the GRP format and illustrate how its components work together, the following walkthrough describes the intended user experience in a hypothetical scenario.

A researcher begins a new project and creates a collection of papers in Zotero. To get an initial orientation, they open the GRP sidebar (Fig. 5.4). The system performs an initial proactive probe of their papers, presenting a set of 'Core' pointers. The anatomy of each pointer is designed for verifiability, containing a title, a description, and interactive citations (Fig. 5.7). The generation process often produces a loose hierarchy of these pointers: a broad, high-level theme common to all papers might appear first, followed by more specific pointers connecting a smaller subset of documents. These pointers may correspond to either explicit knowledge blocks (such as a shared methodological feature) or less explicit ones (such as a recurring concept). Clicking on the generated citations leads the researcher to the corresponding section of the text in a pop-up reader, with the relevant passage highlighted (Fig. 5.5).

As they delve into their work, they read a paper and, using Zotero's native reader, highlight a paragraph they identify as a key knowledge block. Later, they add a new paper with a novel methodology to the collection. These natural actions, in turn, trigger the system's reactive probes, which are announced by banners in the UI inviting the user to generate new, contextualized suggestions (Fig. 5.8).

To explore how this new paper fits into the collection, perhaps to validate their own initial hypothesis about its connections, the researcher generates a new set of pointers. A new pointer marked ‘From New File’ appears, identifying that while most papers in the collection use method X, this new paper uses method Y, providing direct citations to the relevant methodology sections in each. Another new pointer, marked ‘From Highlights’, connects the specific detail they highlighted to a related concept in another paper. Examples of these reactive pointers, differentiated by their colored tags, can be seen in Fig. 5.6.

To synthesize this new connection, the researcher writes a short note in Zotero, linking method Y to the established use of method X. This action triggers another reactive probe, and the resulting ‘From Note’ pointer matches the direction of their thinking, recording and grounding their hypothesis with verifiable links back to the relevant sections. This evolving loop—where the user’s natural interactions with their knowledge base generate context for subsequent, more personalized pointers that aid navigation—is central to the GRP format.

```
[system prompt]
```

Added to further Instructions for Interactive Generation:

```
* A list of previously generated themes is provided for
context.
* Your response should consist of new suggestions derived
from the provided user context (highlights, notes,
etc.), or updated versions of the previous themes if
the new context is directly relevant to them
* Each suggestion must be grounded in and make specific
reference to the provided user context.
* Every suggestion must begin with either the "New:"
prefix for novel themes or the "Update:" prefix for
revisions.
* Do not generate any items that do not adhere to this
format, and keep the total number of suggestions to a
minimum.
```

Added to the following content and data blocks:

```
Previous analysis:
'{kept_headers}'
```

Plus one of the following:

```
Highlights
'{unused_highlights}'
```

or

```
Researcher notes:
'{unused_notes}'
```

or

```
Newly added files:
'{new_files_content}'
```

Figure 5.9: The additional data blocks added to the system promtp for reactive generation, incorporating the system promtp, as welll as additional instructions and the placeholder variables for the Zotero data.

6
Evaluation

Having explored researchers' current workflows to answer our first research sub-question (Section 4), and having detailed the Generative Retrieval Probing (GRP) system as a proposed answer to our second – *What design principles and interaction format can position an LLM as a reflective tool, shifting its outputs from full answers to outputs that invite further inquiry?* (Section 5) – in this second study we shift from design to evaluation. Its purpose is to answer our third and final research sub-question: *How do researchers interact with and perceive a reflective AI tool, and what is the possible effect of this interaction on their cognitive processes and established workflows?*

6.1
Evaluation planning

Evaluating the effectiveness of a system designed to support a researcher's thinking process presents unique challenges. A straightforward measure of success, such as whether participants find the system's recommendations correct, would not capture the potential for such a tool to have a deeper impact through integration with a researcher's ongoing work. Our evaluation, therefore, required a methodology that could gauge not only the tool's utility but also how well its suggestions align with what participants might propose themselves, and how it might facilitate the reflective thinking identified as crucial in our background review.

To that end, we opted to conduct a qualitative study employing the think-aloud method (Lewis, 1982), allowing us to observe in real time how participants engage with the system. However, incorporating the dimension of alignment between the tool's suggestions and the participant's thought process required a more specialized plan. This led us to develop a two-tiered approach: a first session was dedicated to discussing research topics with participants to understand their work, followed by a second session where they interacted with the tool using their own curated document collections. A fundamental aspect of this evaluation is the integration of the tool with the researchers' existing knowledge bases. This design choice meant that researchers served not only as participants but also as the sources of the data used in the study, ensuring that the evaluation was grounded in their authentic work and expertise.

Furthermore, given that our interests for this work encompassed both a researcher's process when working with familiar texts and their strategies when exploring new topics, the study was designed to create use cases for both scenarios. As detailed in the study script, we prompted participants to create collections of papers from topics they knew well alongside a topic they were less familiar with, allowing us to observe how domain knowledge influences their interaction with and assessment of the tool's suggestions.

6.1.1 Goals

The primary goal of this study was to investigate how researchers would use the tool we are proposing and how its use can relate to their existing research practices. Given that the tool's core functionality revolves around generating suggestions based on user input, we again broke out goal down into four more questions that guided our investigation, from perception and sense-making to the potential for workflow integration:

3.1 How do researchers perceive and make sense of AI-generated thematic connections across multiple documents?

3.2 How does their familiarity with a topic influence their interaction with and assessment of the suggestions?

3.3 How do explicit user actions, such as adding new documents or highlighting text, shape the system's subsequent recommendations?

3.4 In what ways, if any, could such a tool be integrated into their existing research practices and workflows?

6.1.2 Procedure and materials

Participant recruitment involved reaching out to both participants from our preliminary study and new participants. We decided on this approach to allow us to follow up with researchers whose work context we already understood, while also diversifying our participant base with individuals who had no prior contact with the project. For returning participants, a screening form was used simply to confirm availability. For new recruits, the form collected the same background information as in the first study (e.g., academic career duration, highest degree, research context) to ensure a consistent set of demographic data across our work. Our final pool for this study was 8 participants, evenly split between 4 returning participants and 4 new ones. Table 6.1 shows the final distribution of participant profiles accross experience and latest academic degree, further separated by purely academic experience, and both academia and industry; as well as the distribution of these participants along the different fields. Returning participants are marked in grey background.

Table 6.1: Final distribution of participant profiles accross experience, latest academic degree, area, and field.

Participant & Area	Academic Status	Exp. (Yrs)
Industry Participants		
Anthropology/Computer Science		
P6 - AI	Doctorate complete	20
Computer Science		
P7 - Explainable AI	Doctorate in progress	8
Engineering		
P8 - Energy	Master's complete	4
Academia Participants		
Computer Science		
P3 - Human-Computer Int. (HCI)	PhD in progress	7
P4 - Software Engineering	Master's complete	5
P5 - Graphs	Master's in progress	4
Engineering		
P1 - Civil Engineering	PhD in progress	10
Physics		
P2 - Particles	PhD in progress	4

▨ Returning participant

Session 1: Curation of Document Collections The initial session was a 15-20 minute semi-structured interview designed to create personalized document collections. This step was crucial for grounding the evaluation in the participant's own work. To explore how familiarity with a topic might affect interaction, we asked participants to identify two or three distinct sub-themes from their research. For two of these themes, with which they were most familiar, they provided 3-5 representative papers. For a third theme, one with which they had less direct experience, they provided at least one starting reference article, which we used to find additional relevant texts. This process resulted in a set of small, thematically coherent collections tailored to each participant's expertise and designed to elicit different interaction stances.

Session 2: Think-Aloud Evaluation The second session was a 60-minute evaluation using a think-aloud protocol. Participants were guided through installing the prototype browser extension and then proceeded through three phases:

1. **Initial Exploration**: Participants began by exploring the suggestions pre-generated for their collections. They were instructed to first survey all suggestion cards for a given collection before selecting one to investigate more deeply. While exploring, they narrated their thoughts, commenting on the relevance of the suggested themes in relation to their own understanding of the collection. Upon choosing a specific suggestion, participants interacted with the linked text passages, clicking on them to be taken to the specific location in the source PDF. They then assessed whether the highlighted passage, figure, or equation adequately supported the tool's claim. This process was repeated for multiple collections.

2. **Interactive Generation**: This phase tested the tool's dynamic, reactive capabilities, a core component of the GRP proposal. Participants were guided through three specific actions outlined in the script: adding a new article to a collection, highlighting passages of interest within an existing article, and writing a collection-level note with a research question. After each action, they triggered the generation of new suggestions and evaluated their relevance and quality, allowing us to observe how they interpreted the system's response to their direct input.

3. **Reflective Debriefing**: The session concluded with a semi-structured interview. Participants were asked to reflect on the tool's overall alignment with their research focus, the utility of the connections it proposed, and its potential fit within their workflow.

6.2 Analysis

All sessions were screen- and audio-recorded, as for the second study it was necessary to store the video of the participant's interaction with the tool. In the same manner as the prior study, they were subsequently transcribed for qualitative analysis, followed by a pass to anonymize them by removing any specific mentions of individuals' names or institutions. We then once again employed thematic coding on the transcriptions. We opted not to adhere to a strict coding methodology, instead focusing on selecting passages that shared common themes and messages that stood out as noteworthy and tied back to our research topics and core goals.

Our rolling codebook for this analysis was composed of 25 codes at the conclusion of this step. To connect the findings of our two studies, the high-level codes from the first study's analysis (as detailed in section 4.3) were used as an initial analytical lens, allowing us to examine how previously identified behaviors and perceptions appeared during interaction with the prototype. We then moved to affinity mapping of all 31 codes (the 25 new codes and the 6 high-level codes from the prior study), grouping them based on shared themes and similarities, which resulted in the final 7 thematic discussion sections presented below. In the following discussion of our findings, specific quotes are numbered when relevant, which can be found in Appendix C.1.

6.3 Discussion

6.3.1 Prototype and LLM limitations

Before we proceed with more complex themes emerging from our analysis, it is productive to first discuss some problems, limitations, and challenges we faced with the prototype. These range from user interface flaws to general issues with how the model generates and supports its suggestions. While some problems are attributable more to the prototype nature of the tool, others point to interesting challenges in designing LLM-based systems for knowledge work, some of which are also bases for themes we discuss with more detail in the further sections. On the smaller issues we faced surrounding the implementation, most were simple UI bugs and workflow failures that created friction in the user experience. For instance, participants noted cases where generated suggestion cards appeared broken or incorrectly rendered (P3), where the interface failed to provide clear feedback on background processes (P8), and, in a few cases, where PDF pages rendered out of order within the tool's document viewer (P1, P6). In three of the sessions (P1, P3, P5), one of the interactive generation features failed, and only in one of those (P5) we had enough time for a successful retry. More significantly, **the most frequent problem participants encountered was with the citations generated by the LLM to support its claims**. This finding resonates with our preliminary study, where participants emphasized the importance of understanding the reason for a citation. While that context referred to formal academic citations, the principle extends here; when the LLM's reasoning for a citation is opaque or flawed, the connection is lost. Participants commonly expressed that while the high-level themes were often accurate, the supporting evidence was the primary point of failure, with one participant summarizing the sentiment as, "The general theme is great, the problem is the citation" [Q1] (P4). In these cases, the unreliability of the supporting passages led to distrust in the feature. One researcher (P4), upon finding a series of unhelpful citations, concluded they would abandon the feature and revert to a manual keyword search (by itself a common behavior to use alongside the citations, which we detail later) to find the evidence themselves, demonstrating how such failures can push users away from the tool and back toward established, more reliable methods [Q2]. **This problem had a particular identity when the system pointed to sections of a paper that were too broad to be useful**. In several instances, participants (P3, P5) found that citations pointing to titles, keywords, or abstracts were not helpful. One participant (P3) dismissed a citation to a title as being "bad, it's just a title" [Q3], while another (P5) noted that a citation to an abstract provided only "a first approach, but perhaps with information that is not so detailed" [Q4]. In these instances, the citation was not technically incorrect but failed to provide the specific information a researcher would look for. However, **the most frequent and significant form of this issue arose in STEM-focused papers where the LLM cited dense, symbolic artifacts like equations, figures, or tables as the sole evidence for a claim**. This practice placed a significant cognitive burden on participants (P1, P7), who were forced to interpret the artifact without surrounding textual explanation. As one researcher noted, it was "very complicated to draw that conclusion from that snippet of an equation" [Q5] (P1), while another found a citation to model parameters to be "isolated and useless" [Q6] (P7). This behavior represents a failure of the system as an epistemic tool. The failure is not just a lack of context, but an inability to parse the internal grammar of these non-prose artifacts; one participant (P6) observed that the model could likely read the text in a diagram but could not interpret the "interpretive dimension" of the "arrows... this big circle" [Q7]. A similar failure was noted (P5) when the model quoted a passage but removed the italics from key terms, losing a "small semantic layer" of emphasis crucial to the academic text [Q8]. Finally, though noticeably rare in our study, **there were instances where the LLM was completely wrong or hallucinated factually incorrect information**. In one case, the model generalized about ethical considerations in a context where a participant (P4) felt they did not apply, concluding, "it seems to me that it hallucinated something" [Q9]. These moments often led to the participant authoritatively correcting the model's error. For instance, when the tool reversed the relationship between experimental and numerical validation, a researcher (P1) simply stated, "No, in fact it's the other way around... the experimental is the real one" [Q10]. These moments are critical reminders of the core trustworthiness challenges that persist even with retrieval-augmented systems. Moreover, apart from being completely wrong, it was more common for the LLM results to be too vague, an important tradeoff given our design considerations. We address this in greater detail in the sections below.

6.3.2
Generality vs specificity and over-generalization

Some of the most interesting tensions that emerged from the study revolve around our deliberate design choice to make the tool's suggestions general rather than specific. Rather than delivering a final, self-contained answer, a suggestion was intended to function as a starting point, inviting the researcher to investigate the underlying texts to verify and interpret the proposed connection – pointers to guide the user toward the source material – in the hopes of prompting a process of critical and reflective engagement. **In many moments, this approach functioned exactly as intended**. Participants recognized that the suggestions were high-level and required their own interpretation to become truly useful. One researcher (P7), reflecting on a generated insight, commented that a given passage "supports what's here, but you need to know other things to believe and understand that" [Q1]. This dynamic, where the tool provides a general direction that the user must then refine, connects back to our preliminary study's finding that LLMs can be useful in the same way as an intern: providing a conversation that is productive precisely because it invites correction and deeper engagement. However, **the value of this intended vagueness was highly dependent on the user's stated goals at various moments of the tests, creating a central trade-off**. For a researcher with a focused task, generality could function as a useful filter. One participant (P5), for instance, appreciated that a high-level summary allowed him to initially bypass less relevant sections of a paper, like the literature review, and navigate directly to the core methodology he was interested in. In this context, the tool's generality successfully supported an expert's workflow. Conversely, when the user's goal was to learn about an unfamiliar topic, this same generality was often perceived as a failure to be informative. In some moments, participants expressed that the suggestions were too shallow, with one (P7) stating that the tool "failed miserably at summarizing the papers. It told me nothing I couldn't have guessed from the title" [Q2]. This frustration was connected to a desire for more specific information, with one researcher (P3) articulating a wish for the tool to be "a little more informative" [Q3], such as providing concrete examples rather than just a high-level theme. **This overall tendency toward generality also led to a problematic behavior: over-generalization of concepts**.

This occurred when the model incorrectly extrapolated a specific concept and applied it too broadly, creating connections that were superficial or false. For example, the tool might claim that "all documents" in a collection used a particular methodology when, in fact, only one or two did (P1) [Q4]. In another instance, a participant (P4) dismissed a suggested theme about "iterative processes" as being too generic to be meaningful, since "any neural network is trained that way" [Q5]. A more subtle form of this issue arose when the tool created a plausible but misleading connection that required significant expert knowledge to disentangle. In one such case (P4), the tool correctly identified a technical concept in a paper but failed to grasp its primary context, linking it to another paper in a way that, while not false, caused initial confusion and required the participant to pause and mentally reconstruct the methods' lineage to verify the claim. These instances highlight the challenge of managing abstraction, demonstrating how intentional vagueness can lead to flawed connections that place the burden of verification squarely on the researcher. This is discussed in more detail in the following section.

6.3.3 GRP as a support for reflective thinking

In contrast to the limitations discussed in the preceding sections, the study also revealed numerous instances where the system successfully supported the researchers' thinking process. These cases aligned with the tool's design goals of assisting participants in understanding and reflecting on connections within their literature. The system demonstrated a capacity to function as an aid, allowing for the information retrieval to support comprehension, synthesis, and navigation. In its most direct application, **the tool was effective at providing participants with an expedited understanding of their collected papers**. All participants reported that the system's ability to summarize core ideas, when working well, was useful, serving as a method for gaining a brief overview or as a reminder of a document's key contributions. As one researcher stated, "they are giving me a summary of what is being discussed, so I can get a summarized view here without having to open the whole text" [Q1] (P1). This support for comprehension also extended to unfamiliar topics; one participant (P3) noted that when exploring a new field, the tool's topic clusters provided a valuable "direction for reading," serving as an initial scaffold [Q2]. **The tool also functioned as a navigational aid, helping researchers move more efficiently within and between texts**. The interactive citations, designed for verifiability, were used by participants to assist their workflow. As one researcher (P6) described, the feature was "a way for me to navigate internally in the text... I navigated more quickly" [Q3]. By linking a high-level theme to a specific passage, the citations effectively reduced the cognitive load required to find a relevant starting point for further exploration. The main point of this functionality was to allow users to more directly navigate to, and between, knowledge blocks like we observed in the first study, which we consider to have been achieved. As part of this, **a nearly universal behavior observed across the study was the use of the tool's suggestions to scaffold manual search and verification**. The themes and keywords presented in the suggestion cards provided participants with specific terms and concepts to look for within the source documents. This often manifested as an explicit desire to use the search function inside the PDFs (by using Ctrl + F) to validate a proposed connection. As one participant (P6) reflected, the impulse was automatic: upon seeing a theme, his first thought was to search for those keywords inside the text, driven by a "desire to navigate with Ctrl+F" [Q4]. This shows the tool's suggestions functioning as prompts that guided and aided the participants' own investigative processes. Beyond summarizing individual points, **the system demonstrated a capacity to help participants identify and structure relationships between them**. This was facilitated by the design of the generated suggestions, which often presented a broad theme followed by more specific sub-themes. We observed this structuring at multiple scales. In some instances, the tool correctly identified broad themes that were common to all texts in a collection (P1) [Q5]. More frequently, it excelled at a more nuanced task, with several participants (P4, P5) noting its ability to "slice" a theme by identifying concepts present in only a subset of the articles [Q6] [Q7]. In one case (P8), a participant observed that the tool had correctly segregated suggestions for a collection into the distinct categories of "domain" and "technique," a structure they affirmed was an accurate representation of their own mental model. This capability to externalize and structure connections is particularly relevant given the findings from our preliminary study, where researchers valued understanding the connections between papers but rarely had a formal practice for storing them. The GRP system demonstrated a potential to address this gap by providing an initial structure of these relationships for the researcher to reflect upon, validate, or reject. This process reminded us of another finding from our first study: the value of receiving a curated set of papers from a colleague who has already provided context on their relevance. By providing these structured starting points, the system demonstrated a capacity to operate in a manner consistent with the principles of epistemic tools. It did not perform the conceptual work for the researcher, but instead provided initial thematic threads and connections as material for their own reflective analysis. This approach was articulated by one participant (P6), who suggested that for AI to be helpful in science, developers must focus on "helping the process of people instead of doing the process for them" [Q8]. By functioning as an aid rather than an automated solution, we believe the tool showed a potential to support the development of the complex of interrelated concepts and arguments central to academic research.

Convergent and Divergent Thinking Building on these applications, the interactions can be further analyzed through the lens of convergent and divergent thinking. The study was designed, in part, to explore how prior expertise shaped engagement by having participants interact with two distinct types of document collections: those containing texts with which they were familiar, and those focused on topics that were new to them. While the known/unknown topic variable influenced interactions, convergent and divergent thinking emerged to us as a separate framework for understanding the observed cognitive patterns. Of course, a participant's familiarity with the subject matter influenced whether the tool's suggestions prompted convergent or divergent thought, but as we observed, the concepts are not entirely interchangeable. **Convergent thinking occurred when the tool's reasoning aligned with a participant's own understanding**. In these instances, observed in multiple sessions (P1, P4, P6, P8), the system produced a summary or thematic connection that the researcher found to be correct and resonant with their own knowledge, creating a sense of validation. One researcher (P6), for example, noted that the way the tool structured its themes "was very reminiscent of the work I did when I wrote my dissertation" [Q9]. This validation was sometimes expressed with emphasis. One participant (P4), an expert in his collection, described the tool's themes as "super 100% aligned" and remarked that one suggestion had "got the core of the matter" [Q10]. Another (P8), upon seeing the tool structure suggestions in a way that mirrored his own mental model, confirmed it was "exactly what I... imagined would happen" [Q11]. In contrast, **divergent thinking occurred when a suggestion diverged from a participant's line of thought in a way they found productive**. Rather than disagreeing with the suggestion, the researchers in these cases recognized it as a novel idea or connection they would not have made themselves. This was sometimes described as pointing them "not where I would focus, but where I might want to focus" (P5, P8) [Q12]. A clear example of this occurred with a participant (P4) who was exploring an unfamiliar collection. The tool surfaced the term "Representational Learning," which he did not know. He recognized its significance from the context and identified it as a new path for his own learning, stating he could “use it as a basis to better learn the niche” [Q13]. This example illustrates how a single suggestion can provide a new direction for a researcher’s inquiry. **The context of the participant’s expertise played a significant role in mediating these thinking patterns**. When interacting with collections on known topics, participants (P1, P3, P4, P5, P6, P7, P8) could assess the validity of the tool’s suggestions with more confidence. Their expertise allowed them to evaluate both the accuracy of convergent suggestions and the novelty of divergent ones, as they could more easily filter incorrect information and recognize novel connections. Conversely, when working with texts in an unfamiliar domain, the interaction was more exploratory. Suggestions were used more as a starting point for learning (P3, P4), and while participants could still assess the general coherence of the themes, their evaluation was more tentative. In this context, the tool’s ability to provide an initial summary or structure was reported as being helpful for navigating an unfamiliar area.

Misleading Connections In contrast to the productive connections discussed above, **there were also instances where the tool’s attempts to identify relationships between papers resulted in misleading connections**. These occurrences are a direct manifestation of the challenges of hallucination, factual inaccuracy, and over-generalization discussed in preceding sections. They represent a more specific variant of those problems, occurring within the context of the tool’s primary function of synthesizing information to create connections. While this feature was reported as highly valuable when successful, its failures produced connections that were clear errors. These errors were often a direct consequence of the system’s tendency to over-generalize concepts. In some cases, the error was a misrepresentation of a core concept, with one participant (P1) correcting a suggestion by explaining that the tool had reversed the fundamental relationship between experimental and numerical validation in their field [Q14]. In other cases, the tool asserted a thematic link between papers based on a concept that was not present in the source material (P6) [Q15]. At other times, the connections were not just factually incorrect but were perceived as nonsensical or spurious, with participants describing them as instances where the tool was “deceived in a very easy way” (P8) [Q16]. These are flawed attempts at the synthesis that, when successful, participants reported as being highly valuable. Such errors highlight a direct trade-off of using a generative system for this type of task. The ability to produce novel, divergent connections is intrinsically linked to the risk of producing plausible but misleading ones, which reinforces the cognitive burden of verification and the fragile nature of user trust discussed earlier.

6.3.4 GRP features and the reflective experience

Participants used the different interactive features with distinct, goal-oriented intentions, dynamically shifting their approach based on the analytical task at hand. Generating from **highlights** was typically used for more granular, bottom-up exploration; it was a way for users (P3, P5, P7) to signal a specific passage of interest and ask the tool to elaborate or find similar concepts. The **note** feature, in contrast, was used for more top-down, deliberate inquiries. It allowed participants to pose a direct, open-ended question to their collection (P6, P8) or to test a specific hypothesis about a cross-document connection they had already formulated (P4). Finally, generating from **new files** was used by participants (P1, P4, P8) to explore how a collection's thematic structure evolved over time, for instance by checking for consistency with prior themes [Q1] or by exploring what new links might emerge. We observed that **a majority of participants (P3, P4, P5, P6, P8) used these features with clear intent, consciously providing their own context to steer the system's focus**. This was often framed as a way to initiate a dialogue with their literature. One researcher (P6), for example, described using the note feature as a way to "ask a question to the documents" [Q2]. Another (P8) saw the features as a way to "bootstrap" his thinking process, using his own interactions to provide the initial context needed to get the system started [Q3]. In several instances (P3, P5, P8), this user-provided context led to moments of insight. When the system successfully interpreted a user's input, it demonstrated value in this context-aware partnership. For example, a suggestion generated from one participant's (P3) own highlights successfully "expanded a bit on what I had marked" [Q4], showing the system building directly on his provided context. In another case (P5), a suggestion prompted by his highlights sparked a moment of surprise and divergent thinking: "Oh, that's interesting... I hadn't thought about that" [Q5]. However, **the system's ability to act as a responsive partner was entirely dependent on its ability to correctly interpret the user's provided context**, and breakdowns in this interpretation were noted in a few cases (P4, P6, P7). In one case (P4), a participant wrote a note with a clear, goal-oriented context: he wanted the tool to find a conceptual bridge between two specific papers. When the system failed to identify this link, he noted the failure of the tool to understand his intent: "I expected a bridge... and it didn't make a bridge"

[Q6]. In another, more nuanced example (P7), a participant diagnosed the tool's misinterpretation of his highlight's semantic context, hypothesizing that it had fixated on the single keyword *random* while ignoring the surrounding text that gave it meaning [Q7]. These cases show that effective probing is a delicate negotiation, requiring the user to not only provide clear context but also to diagnose and correct the system's interpretation when it fails. Ultimately, the tool's features were designed to work as an integrated whole to support a full cycle of reflective, context-driven inquiry. The interactive generation features allowed the user to steer the analysis and produce a new set of hypotheses based on their own evolving interests. Subsequently, the interactive citations within these new suggestions provided the crucial mechanism for verification. This completed the reflective loop: the user provided context, the tool reacted with a focused suggestion, and the user could then immediately verify and explore the basis for that reaction within the source texts. This integration of user-directed generation and verifiable evidence is what allowed the system, when successful, to fulfill its design purpose as a responsive partner in the user's cognitive process.

6.3.5 Other characteristics of working with LLMs

Beyond the qualities of the generated suggestions themselves, our analysis revealed other interesting characteristics of the participants' interactions with the LLM. These relate to how researchers perceive the system, the mental models they form to explain its opaque behavior, and how their trust in the tool evolves. **A recurring issue, observed in a few sessions (P6, P7), was the model's misinterpretation of rhetorical and structural cues in academic texts**. One participant (P6) noted that the tool had read a transitional phrase, a "linguistic resource", and interpreted it as a substantive point [Q6]. Similarly, another researcher (P7) observed that the tool would frequently cite non-core sections of a paper. This included citing the "Prior Work" section to support a high-level claim [Q7] or pointing to a table of technical parameters, which, without the surrounding context, was "isolated and useless" [Q8]. **This type of unpredictable behavior contributes to a broader challenge: the mental model mismatch between the user and the system**. Because the LLM's inner workings are opaque, in a few cases participants (P6, P7) became preoccupied with trying to diagnose its behavior rather than focusing on its suggestions. This was evident when they began to formulate folk theories to explain the model's flawed reasoning. For instance, one researcher (P6) hypothesized that the tool was biased toward citing sections that were already summaries or introductions, while another (P7) theorized that when the tool "can't fit [a paper] into the narrative, it excludes it" [Q9]. In these moments, the user's role shifted from that of a collaborator to that of a diagnostician, creating a cognitive burden of not only verifying the output but also attempting to reverse-engineer the opaque process that produced it. This uncertainty directly impacts the development of trust, which we observed to be an evolving feeling that shifted throughout the interaction. For some participants (P4, P6, P7), **a single error had a cascading effect, retroactively casting doubt on suggestions that had previously seemed useful**. As one participant (P6) explained, a few incorrect passages were enough to "increase my distrust even of what seemed useful" [Q10]. This fragile nature of trust is a significant factor when considering the use of LLMs as collaborative partners in knowledge work, as the reliability of the entire interaction can be compromised by individual moments of failure. This connects directly to our prior study's findings on untrustworthiness, where participants perceived LLMs not as reliable sources of answers, but as generative partners whose outputs always require careful, critical verification.

6.3.6 Researchers as a demographic for interaction with the tool

Throughout our evaluation, we observed how participants' engagement with the tool was shaped by their training and identity as researchers. Their interactions were not those of a general user seeking information, but of a specific demographic with distinct analytical approaches and established practices. These researcher-specific behaviors influenced their use of the GRP system in several key ways, from how they approached the content of the papers to how they evaluated the tool's place within their work. **A common thread observed across many sessions (P1, P4, P5, P6, P7, P8) was the practice of seeking specific knowledge blocks within the literature**, a key finding from our preliminary study. As observed in our discussion of the tool's navigational support (Section 6.3.4), participants tended not to approach papers as monolithic texts but as containers for discrete pieces of information. This mindset was explicitly articulated by one participant (P5) who, when discussing his goals, stated that for certain tasks "what you want is the method" and that contextual sections like the literature review are "not so interesting" [Q1]. This goal-oriented extraction was a recurring pattern, with researchers focusing on different blocks depending on their needs, from "the procedures" and "the results" (P1) [Q2] to the "proofs, the theorems, not so much on the method" in more theoretical papers (P5) [Q3]. Similarly, we observed that **researchers brought a deeply critical and contextual lens to their interactions**. As discussed in our findings on trustworthiness (Section 6.3.3), participants actively validated the tool's claims against their own extensive domain knowledge, a mindset demonstrated by concerns about missing contradictory information (P6) [Q4] or the tool's failure to recognize the subtle evolutionary relationship between two papers (P7) [Q5]. In general, we observed how information was not taken at face value but instead integrated into existing domain knowledge of their field. This critical lens was applied differently, however, when researchers operated outside their core expertise, revealing two distinct stances toward using the tool for exploration. One was an exploratory stance, where a participant (P4) who was new to a topic productively used the tool as a guide, seeing the novel keywords it surfaced as a "basis to better learn the niche" [Q6]. In contrast, another participant (P8) adopted a skeptic stance, stating a strong reluctance to use the tool for exploratory work precisely because his lack of expertise would leave him unable to critically filter the outputs and determine "what is important, what is not, what is for me, what is not" [Q7]. This shows that a researcher's willingness to engage with the tool for learning is mediated by their confidence in their own critical judgment to verify and contextualize its output. Finally, a noteworthy aspect of the participants' feedback, observed in multiple sessions (P3, P4, P7, P8), was **how they consistently evaluated the GRP system in terms of its potential role within specific stages of their established research workflows**. While our work, like the select few in the related literature, aims to support the general knowledge process rather than automating specific, final tasks, it was natural for this demographic to map the tool's capabilities onto their more concrete activities. Participants identified its potential as an aid for various tasks, such as providing a "quick way to find the detail... and use it in your 'related work' section" (P4) [Q8], or for getting an initial "direction for reading" when beginning a Systematic Literature Review (P3) [Q9]. Others saw its value in specific moments of their workflow, such as for incrementally integrating new literature into a known corpus of work (P8) [Q10] or for situating their own paper within the broader literature just before submission (P7) [Q11].

6.3.7
Moving beyond the (vague) suggestions format

Expanding on some of the points raised in the sections above, we believe the tool's intentionally vague format was sometimes a productive starting point for divergent thinking but was often a source of frustration when it led to unhelpful or misleading connections. This final theme synthesizes participants' feedback on this tension, capturing their desire to move beyond passively receiving suggestions and instead actively steer the tool's analytical process toward their own specific goals. Throughout the study, **participants stated a clear desire for more agency and control over the AI's reasoning**. This was framed not just as a request for better features, but as a guiding principle for how such tools should be designed to support scientific work. One researcher (P6) articulated this by arguing that for AI to be truly helpful, developers must focus on "helping the process of people instead of doing the aprocess for them" [Q1]. He warned that a tool that fails to do so risks encouraging researchers to "outsource their reading," a practice that fosters a deep sense of distrust and undermines the scholarly activity of reading [Q2]. This sentiment reflects the demand for tools that are less like black-box answer providers and more like directable, transparent instruments. This desire for agency was evident in participants' goal to move from passively interpreting suggestions to actively steering the tool's focus. They wanted to provide more explicit context and direction to the AI's analysis. For instance, one researcher (P5) stated that the tool would be most useful "if you can establish your context for the tool. What you want to look at in the papers" [Q3]. Another participant (P4) used the interactive note feature with the stated intent of directing the tool to find if other architectures use similar ideas, demonstrating an intent to use the system to answer a specific, self-generated research question. **One interesting way a participant sought to guide the analysis was looking for a chronological dimension**, a finding observed in multiple sessions (P4, P7). One researcher (P4) noted that his goal was not only to see thematic overlaps but also to construct a "chronology to understand where one builds on the other" [Q4]. Another (P7) independently raised this issue, observing that the tool failed to recognize that two papers were by the same author and one was an evolution of the other; it "doesn't show how the papers relate to each other," only how they relate to a static theme [Q5]. This recurring theme reflects established information-seeking behaviors like forward and backward chaining and suggests an opportunity for the tool to organize connections not just thematically, but also relationally, illustrating the intellectual lineage of concepts within a collection. Looking forward, these interactions point toward several potential avenues for evolving the GRP model beyond its current format of vague suggestions. Participants suggested a tool that could be instructed to perform more specific analytical tasks. This included a "deep dive" mode capable of producing a "meta-analysis" of a given topic (P8) [Q6], as well as more direct writing support, such as allowing a user to write a paragraph in a note and have the tool find supporting citations for it (P8). Others suggested making the suggestions themselves more descriptive; as a direct response to the frustration with over-generalization discussed in Section 6.3.2, one participant (P3) suggested that the cards should list a few of the connections found, rather than just stating a high-level theme.

7
Conclusions and future work

Drawing from the findings of both our preliminary exploration of researcher workflows and our evaluation of the Generative Retrieval Probing prototype, we propose a set of broader conclusions and chart a path for future work. This research began with a central question: how can LLM-based systems move beyond the role of definitive answer providers to become reflective epistemic tools that support the nuanced cognitive processes of academic work? We believe, given the results of our second study, that the GRP interaction format represents a tangible step in this direction.

The importance of pursuing such alternatives to current LLM designs is underscored by the recent neuroscientific findings we mentioned in Section 1. The aforementioned work of Kosmyna et al. provides compelling evidence for the phenomenon of cognitive debt, where reliance on LLMs for complex tasks correlates with diminished neural connectivity and impaired memory encoding. While our qualitative evaluation does not measure brain activity, our proposal aligns with the pressing need for systems that counter the trend toward passive consumption that their work identifies. By re-centering the system on its retrieval aspect, GRP is designed to encourage the very active information seeking that Kosmyna et al. found to be more cognitively engaging. We hope that by presenting user-verifiable suggestions that point said user back to diverse sources for evaluation, our format can leverage an LLM's generative power without promoting excessive cognitive offloading. Our evaluation suggests this is a promising path; participants used the tool's suggestions not as final answers, but as starting points to scaffold their own investigation, supporting in different ways both the convergent exploration of their own ideas and the divergent discovery of novel connections they had not previously considered, with productive exploration following.

The GRP format is therefore presented not as a finished product, but as a foundational framework for future explorations. The broader challenge for designing knowledge tools is to create interaction paradigms that foster a reflective, collaborative partnership between user and machine. As one of our participants (P8) articulated, the ultimate goal must be to create tools that "help the process of people instead of doing the process for them."

Future research must continue to prioritize this human-centered goal, our hopes being that as generative systems become more powerful, they are designed to augment human intellect rather than supplant it.

Chiefly, the evaluation study also brought to light a series of trade-offs inherent in our current design, which we believe bring about the most interesting conclusions we can draw from this initial research. As detailed in the conclusions from our evaluation in Section 6.3, a central challenge is the balance between fostering reflection and addressing issues related to suggestions that were sometimes overly general. These conclusions suggest that the generality of the suggestions, while often a productive starting point, could also introduce a cognitive burden of verification for the user, a trade-off that warrants further exploration.

In connection to these trade-offs, a clear direction for future work involves refining how the system generates effective reflective pointers: prompts that extend beyond generic thematic summaries. Future work could aim to better leverage the system's strengths in supporting a researcher's thinking process while mitigating its weaknesses. The lens of convergent and divergent thinking, which emerged from our evaluation as a useful descriptor of participant interactions, offers one such avenue. Future iterations could explore how to more directly represent these modes of thought within the system, for instance, by allowing a user to orient the suggestions they receive toward themes that align with their existing highlights or toward more novel connections found in less-interacted-with documents. An additional approach involves tagging suggestions by their semantic function (e.g., novel connection, methodological parallel, or core concept summary), which would allow users to filter the system's output based on their immediate exploratory goal.

A key insight derived from our work is the consistently expressed desire of researchers to move beyond passively receiving suggestions and toward actively steering the tool's analytical focus. A minimal form of this would be to allow users to react to suggestions by registering positive or negative feedback or by deleting irrelevant cards to prevent them from influencing future generations. Creating a more useful feedback loop could involve allowing a user to provide completely free-form textual feedback on a suggestion. This feedback could act as a meta-note that informs the system's context for subsequent generations, creating a more direct dialogue. A more conceptually aligned next step would fully integrate the generated suggestions back into the researcher's knowledge base. A suggestion could be accepted by the user, transforming it into a new, editable note within their reference manager and making it a primary artifact in their workflow that can be used as a seed for further probes.

Extrapolating from this, a productive direction for future work involves rethinking the user interface. While the card feed of the current prototype was a deliberate choice for a minimal and integrated initial design, our user testing suggests it may not be the most effective paradigm for deep exploratory work. Future designs should retain the core reactive principle of GRP, which involves the tool responding to a user's live work in their knowledge base, while offering more exploratory avenues for interaction. Future work could explore a portfolio of formats, such as chronological maps that illustrate a topic's evolution (a format participants noted was important) or graph-based visualizations that map not just the papers but also weave in the suggestions themselves as new nodes. A particularly promising general avenue, we believe, is the use of more visual and conceptual mind maps where a user could interact with high-level themes first. In such an interface, the direct manipulation of these visual artifacts could evolve into a more comprehensive form of query to the tool. Actions like deleting a branch of a mind map or promoting a concept could become signals that drive the next cycle of generation.

The rapid development of the LLM field also warrants consideration for future work, as many sophisticated techniques matured while this research was being conducted. A particularly significant development is the development of agentic (Yao et al., 2022) workflows, which represent a shift from single-instruction execution to more autonomous, multi-step problem-solving. In this paradigm, a system interprets a high-level goal, decomposes it into a sequence of tasks, and uses a toolkit of specialized functions, such as external search or data analysis, to execute them. For a future iteration of GRP, adopting such a format could allow the system to handle more complex research objectives that require planning and adaptation, moving well beyond the limitations of a single prompt-and-response interaction.

Within this agentic framework, the specific capability of 'agentic retrieval' (Asai et al., 2023) offers a particularly relevant enhancement. This approach elevates information seeking from a static action to a dynamic strategy. Instead of executing a single search, a retrieval agent critically evaluates the relevance and sufficiency of the information it finds. Based on this self-critique, it can autonomously refine its approach by generating new queries or exploring different angles. Integrating this process into GRP could help produce higher-quality context, directly addressing the challenge of overly general or vague suggestions by ensuring the underlying information is more robust and precisely targeted.

With a foundation of agentic workflows and robust retrieval, the system could begin to deliver the deep research capabilities that participants desired.

This agentic workflow for complex information retrieval involves letting the LLM generate plan steps to access information resources by itself, allowing it to, in a way, follow investigative threads in documents independently (Madaan et al., 2023). It is here that the system could trace the evolution of an idea, identify novel connections, or, as participant P3 suggested, generate suggestions that list the specific connections found in detail rather than just a high-level theme. This represents the ultimate evolution of the GRP concept: transforming the tool from a reactive prompter into a true analytical partner capable of contributing to the critical process of knowledge creation. Our own findings revealed a clear desire among participants for the system to evolve beyond reactive suggestions toward more comprehensive, multi-step investigations. As noted in our evaluation (6.3.7), participants envisioned a tool that could perform a deep dive or meta-analysis of a topic or find supporting citations for a claim.

Building on these ideas, a further extension of the GRP concept could involve aiding more directly in the information-seeking process itself. The same generative probing structure could be leveraged to help researchers navigate the iterative process of finding new literature. Drawing from our preliminary study's findings on iterative query building, the system could assist the researcher in building queries for research libraries, analyzing their interactions with existing papers to propose and refine search strategies. The principles of GRP could also be extended beyond academic work to other knowledge-intensive domains where professionals must synthesize and critically engage with large, specialized document collections.

In conclusion, this research set out to reimagine the role of LLMs in the academic workflow, responding to the need for tools that support reflection rather than simply providing answers. Through the design and evaluation of the Generative Retrieval Probing format, we have explored an initial approach that suggests a viable alternative to the prevailing paradigm of question-answering systems. While preliminary, this work proposes a foundational experimental framework and a set of principles for designing tools that function as collaborative partners in the reflective and critical process of knowledge creation.

Bibliography

Abd-alrazaq, A., Nashwan, A. J., Shah, Z., Abujaber, A., Alhuwail, D., Schneider, J., AlSaad, R., Ali, H., Alomoush, W., Ahmed, A., and Aziz, S. (2024). Machine Learning–Based Approach for Identifying Research Gaps: COVID-19 as a Case Study. *JMIR Formative Research*, 8:e49411.

Agarwal, S., Laradji, I. H., Charlin, L., and Pal, C. (2024). LitLLM: A Toolkit for Scientific Literature Review. arXiv:2402.01788 [cs].

Al Ghadban, Y., Lu, H. Y., Adavi, U., Sharma, A., Gara, S., Das, N., Kumar, B., John, R., Devarsetty, P., and Hirst, J. E. (2023). Transforming Healthcare Education: Harnessing Large Language Models for Frontline Health Worker Capacity Building using Retrieval-Augmented Generation.

Asai, A., Wu, Z., Wang, Y., Sil, A., and Hajishirzi, H. (2023). Self-rag: Learning to retrieve, generate, and critique through self-reflection. *arXiv preprint arXiv:2310.11511*.

BAAI (2024). Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through multi-objective optimization. *arXiv preprint arXiv:2402.03216*.

Beel, J., Gipp, B., Langer, S., and Breitinger, C. (2016). Research-paper recommender systems: A literature survey. *International Journal on Digital Libraries*, 17(4):305–338.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., and Liang, P. (2022). On the Opportunities and Risks of Foundation Models.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language Models are Few-Shot Learners. https://arxiv.org/abs/2005.14165v4.

Caramancion, K. M. (2024). Large Language Models vs. Search Engines: Evaluating User Preferences Across Varied Information Retrieval Scenarios.

Cetina, K. K. (1991). Epistemic Cultures: Forms of Reason in Science. *History of Political Economy*, 23(1):105–122.

Chen, Z., Zhang, L., Weng, F., Pan, L., and Lan, Z. (2024). Tailored Visions: Enhancing Text-to-Image Generation with Personalized Prompt Rewriting. arXiv:2310.08129 [cs].

Christakopoulou, K., Lalama, A., Adams, C., Qu, I., Amir, Y., Chucri, S., Vollucci, P., Soldo, F., Bseiso, D., Scodel, S., Dixon, L., Chi, E. H., and Chen, M. (2023). Large Language Models for User Interest Journeys. arXiv:2305.15498 [cs].

Conrad, L. Y., Bruce, C. S., and Tucker, V. M. (2020). Constructing information experience: A grounded theory portrait of academic information management. *Aslib Journal of Information Management*, 72(4):653–670.

Corbin, J. M. and Strauss, A. (1990). Grounded theory research: Procedures, canons, and evaluative criteria. *Qualitative sociology*, 13(1):3–21.

Deldjoo, Y., He, Z., McAuley, J., Korikov, A., Sanner, S., Ramisa, A., Vidal, R., Sathiamoorthy, M., Kasirzadeh, A., and Milano, S. (2024). A review of modern recommender systems using generative models (gen-recsys).

Eger, S., Cao, Y., D'Souza, J., Geiger, A., Greisinger, C., Gross, S., Hou, Y., Krenn, B., Lauscher, A., Li, Y., Lin, C., Moosavi, N. S., Zhao, W., and Miller, T. (2025). Transforming Science with Large Language Models: A Survey on AI-assisted Scientific Discovery, Experimentation, Content Generation, and Evaluation. arXiv:2502.05151 [cs].

Fang, Y., Thomas, S. W., and Zhu, X. (2024). HGOT: Hierarchical Graph of Thoughts for Retrieval-Augmented In-Context Learning in Factuality Evaluation. arXiv:2402.09390 [cs].

Ferrag, M. A., Alwahedi, F., Battah, A., Cherif, B., Mechri, A., and Tihanyi, N. (2024). Generative AI and Large Language Models for Cyber Security: All Insights You Need.

Fu, A., Chan, C.-M., Liu, Y., Zhang, P.-F., Zuo, J.-C., Li, C.-X., Liu, J.-W., Li, W.-Z., Yin, X.-C., and Liu, C.-L. (2024). RAGFlow: A large language model-based deep document understanding framework. *arXiv preprint arXiv:2405.18434*.

Gao, L., Ma, X., Lin, J., and Callan, J. (2022). Precise Zero-Shot Dense Retrieval without Relevance Labels. https://arxiv.org/abs/2212.10496v1.

Gemma Team (2024). Gemma: Open models based on gemini research and technology. Technical report, Google.

Girard, J. and Girard, J. (2015). Defining knowledge management: Toward an applied compendium. 3(1).

Gottweis, J., Weng, W.-H., Daryin, A., Tu, T., Palepu, A., Sirkovic, P., Myaskovsky, A., Weissenberger, F., Rong, K., Tanno, R., Saab, K., Popovici, D., Blum, J., Zhang, F., Chou, K., Hassidim, A., Gokturk, B., Vahdat, A., Kohli, P., Matias, Y., Carroll, A., Kulkarni, K., Tomasev, N., Guan, Y., Dhillon, V., Vaishnav, E. D., Lee, B., Costa, T. R. D., Penadés, J. R., Peltz, G., Xu, Y., Pawlosky, A., Karthikesalingam, A., and Natarajan, V. (2025). Towards an ai co-scientist.

Hoeber, O., Patel, D., and Storie, D. (2019). A Study of Academic Search Scenarios and Information Seeking Behaviour. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, CHIIR '19, pages 231–235, New York, NY, USA. Association for Computing Machinery.

Hoeber, O. and Storie, D. (2022). Information seeking within academic digital libraries: A survey of graduate student search strategies. *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, pages 1–5.

Hu, X., Fu, H., Wang, J., Wang, Y., Li, Z., Xu, R., Lu, Y., Jin, Y., Pan, L., and Lan, Z. (2024). Nova: An Iterative Planning and Search Approach to Enhance Novelty and Diversity of LLM Generated Ideas. arXiv:2410.14255 [cs].

Huang, C., Wu, Z., Hu, Y., and Wang, W. (2024). Training Language Models to Generate Text with Citations via Fine-grained Rewards. arXiv:2402.04315 [cs].

Kathleen Kern, M. and Hensley, M. K. (2011). Citation management software: Features and futures. *Reference and User Services Quarterly*, 50(3):204–208.

Kosmyna, N., Hauptmann, E., Yuan, Y. T., Situ, J., Liao, X.-H., Beresnitzky, A. V., Braunstein, I., and Maes, P. (2025). Your Brain on ChatGPT: Accumulation of Cognitive Debt when Using an AI Assistant for Essay Writing Task. arXiv:2506.08872 [cs].

Lewis, C. (1982). Using the'thinking-aloud'method in cognitive interface design. *Research Report RC9265, IBM TJ Watson Research Center*.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. (2021). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.

Li, L., Xu, W., Guo, J., Zhao, R., Li, X., Yuan, Y., Zhang, B., Jiang, Y., Xin, Y., Dang, R., Zhao, D., Rong, Y., Feng, T., and Bing, L. (2024a). Chain of Ideas: Revolutionizing Research Via Novel Idea Development with LLM Agents. arXiv:2410.13185 [cs].

Li, X., Jin, J., Zhou, Y., Zhang, Y., Zhang, P., Zhu, Y., and Dou, Z. (2025). From Matching to Generation: A Survey on Generative Information Retrieval. arXiv:2404.14851 [cs].

Li, X. and Ouyang, J. (2024). Explaining Relationships Among Research Papers. arXiv:2402.13426 [cs].

Li, X., Zhu, C., Li, L., Yin, Z., Sun, T., and Qiu, X. (2024b). LLatrieval: LLM-Verified Retrieval for Verifiable Generation. arXiv:2311.07838 [cs].

Li, Y., Chen, L., Liu, A., Yu, K., and Wen, L. (2024c). ChatCite: LLM Agent with Human Workflow Guidance for Comparative Literature Summary. arXiv:2403.02574 [cs].

Li, Z. and Zou, X. (2019). A Review on Personalized Academic Paper Recommendation. *Computer and Information Science*, 12(1):p33.

Liu, H., Zhou, Y., Li, M., Yuan, C., and Tan, C. (2025). Literature Meets Data: A Synergistic Approach to Hypothesis Generation. arXiv:2410.17309 [cs].

Liu, J., Liu, C., Zhou, P., Lv, R., Zhou, K., and Zhang, Y. (2023). Is ChatGPT a Good Recommender? A Preliminary Study. arXiv:2304.10149 [cs].

Liu, Y., Chen, S., Cheng, H., Yu, M., Ran, X., Mo, A., Tang, Y., and Huang, Y. (2024). CoQuest: Exploring Research Question Co-Creation with an LLM-based Agent. arXiv:2310.06155 [cs].

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., and Ha, D. (2024). The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. arXiv:2408.06292 [cs].

Luhmann, N. (1981). Kommunikation mit zettelkästen: Ein erfahrungsbericht. In *Öffentliche Meinung und sozialer Wandel/Public Opinion and Social Change*, pages 222–228. Springer.

Madaan, A., Tandon, N., Gupta, P., Hall, K., Gao, L., Majumder, S., McAuley, J., Narayan, S., Oh, J., Precup, D., et al. (2023). Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Markauskaite, L. and Goodyear, P. (2017). Epistemic Tools and Artefacts in Epistemic Practices and Systems. In Markauskaite, L. and Goodyear, P., editors, *Epistemic Fluency and Professional Education: Innovation, Knowledgeable Action and Actionable Knowledge*, pages 233–264. Springer Netherlands, Dordrecht.

Meyer, J. and Zimmermann, C. D. (2022). Unlocking information from unstructured documents with 'deepdoctection'. *arXiv preprint arXiv:2209.11305*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space.

Mizrachi, D. and Bates, M. J. (2013). Undergraduates' personal academic information management and the consideration of time and task-urgency.

Journal of the American Society for Information Science and Technology, 64(8):1590–1607.

Muennighoff, N., Tazi, N., Magne, L., and Reimers, N. (2022). MTEB: Massive text embedding benchmark. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6154–6186. Association for Computational Linguistics.

Mysore, S., Lu, Z., Wan, M., Yang, L., Sarrafzadeh, B., Menezes, S., Baghaee, T., Gonzalez, E. B., Neville, J., and Safavi, T. (2024). Pearl: Personalizing large language model writing assistants with generation-calibrated retrievers.

Nichol, A. J., Hastings, C., and Elder-Vass, D. (2023). Putting philosophy to work: Developing the conceptual architecture of research projects. *Journal of Critical Realism*, 22(3):364–383.

Osae Otopah, F. and Dadzie, P. (2013). Personal information management practices of students and its implications for library services. In *Aslib Proceedings*, volume 65, pages 143–160. Emerald Group Publishing Limited.

Pu, K., Feng, K. J. K., Grossman, T., Hope, T., Mishra, B. D., Latzke, M., Bragg, J., Chang, J. C., and Siangliulue, P. (2024). IdeaSynth: Iterative Research Idea Development Through Evolving and Composing Idea Facets with Literature-Grounded Feedback. arXiv:2410.04025 [cs].

Radensky, M., Shahid, S., Fok, R., Siangliulue, P., Hope, T., and Weld, D. S. (2025). Scideator: Human-LLM Scientific Idea Generation Grounded in Research-Paper Facet Recombination. arXiv:2409.14634 [cs].

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training.

Ritala, P., Ruokonen, M., and Ramaul, L. (2024). Transforming boundaries: How does ChatGPT change knowledge work? *Journal of Business Strategy*, 45(3):214–220.

Salemi, A., Mysore, S., Bendersky, M., and Zamani, H. (2024). LaMP: When Large Language Models Meet Personalization. arXiv:2304.11406 [cs].

Scherbakov, D., Hubig, N., Jansari, V., Bakumenko, A., and Lenert, L. A. (2025). The emergence of large language models as tools in literature reviews: a large language model-assisted systematic review. *Journal of the American Medical Informatics Association*, 32(6):1071–1086.

Susnjak, T. (2023). PRISMA-DFLLM: An Extension of PRISMA for Systematic Literature Reviews using Domain-specific Finetuned Large Language Models. arXiv:2306.14905 [cs].

Tamminen, K. A. and Poucher, Z. A. (2008). Research philosophies | 39 | The Routledge International Encyclopedia.

Vakkari, P. (2001). (PDF) A theory of the task-based information retrieval process: A summary and generalisation of a longitudinal study.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention Is All You Need.

Wang, D., Huang, Q., Jackson, M., and Gao, J. (2024a). Retrieve What You Need: A Mutual Learning Framework for Open-domain Question Answering. *Transactions of the Association for Computational Linguistics*, 12:247–263.

Wang, W., Lin, X., Feng, F., He, X., and Chua, T.-S. (2024b). Generative Recommendation: Towards Next-generation Recommender Paradigm. arXiv:2304.03516 [cs].

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., and Zhou, D. (2023). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.

Xu, S., Pang, L., Shen, H., Cheng, X., and Chua, T.-S. (2024). Search-in-the-Chain: Interactively Enhancing Large Language Models with Search for Knowledge-intensive Tasks. arXiv:2304.14732 [cs].

Yang, Z., Liu, W., Gao, B., Xie, T., Li, Y., Ouyang, W., Poria, S., Cambria, E., and Zhou, D. (2025). MOOSE-Chem: Large Language Models for Rediscovering Unseen Chemistry Scientific Hypotheses. arXiv:2410.07076 [cs].

Yao, S., Zhao, J., Yu, D., Du, N., Durmus, E., Lin, T., German, O., Pinto, L., Dosovitskiy, A., and Gnanamanickam, A. (2022). ReAct: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629.*

Ye, X., Sun, R., Arik, S. , and Pfister, T. (2024). Effective Large Language Model Adaptation for Improved Grounding and Citation Generation. arXiv:2311.09533 [cs].

Zeleny, M. (1987). Management support systems: Towards integrated knowledge management. *Human Systems Management*, 7(1):59–70. Publisher: SAGE Publications.

Zhao, Y., Singh, P., Bhathena, H., Ramos, B., Joshi, A., Gadiyaram, S., and Sharma, S. (2024). Optimizing LLM Based Retrieval Augmented Generation Pipelines in the Financial Domain. In Yang, Y., Davani, A., Sil, A., and Kumar, A., editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 279–294, Mexico City, Mexico. Association for Computational Linguistics.

Zhiyuli, A., Chen, Y., Zhang, X., and Liang, X. (2023). BookGPT: A General Framework for Book Recommendation Empowered by Large Language Model. arXiv:2305.15673 [cs].

Zhou, Y., Liu, H., Srivastava, T., Mei, H., and Tan, C. (2024a). Hypothesis Generation with Large Language Models. pages 117–139. arXiv:2404.04326 [cs].

Zhou, Y., Zhu, Q., Jin, J., and Dou, Z. (2024b). Cognitive Personalized Search Integrating Large Language Models with an Efficient Memory Mechanism. arXiv:2402.10548 [cs].

A
Preliminary study recruitment form

Translated from Portuguese

- How long have you been working in academic research?
  - *(Open-ended)*
- What is your highest completed degree?
  - Undergraduate
  - Specialization
  - Master's
  - Doctorate
- In what context do you conduct research?
  - Academic institution
  - Industry
  - Other... *(Open-ended)*
- How often do you conduct research alone or in collaboration with others?
  - *(Scale: 1 - Never to 5 - Always)*
  - Alone
  - In collaboration with others
- What is your primary research area(s)?
  - *(Open-ended)*
- Would you be interested in participating in a 45-90 minute conversation for this study?
  - Yes
  - No
  - It depends! I'd like to know more
- What is your name?
  - *(Open-ended)*
- What is your contact information?
  - *(Open-ended)*

B
Preliminary study interview script

Translated from Portuguese

- Which tool or tools do you use to find papers related to your research topics? (If the answer is more related to systematic reviews, ask about other cases)
- Which tool do you use the most and why?
- Do you feel that among these tools there is one that helps you find papers that are more useful to you?
- Is it the same one you use the most?

- Among the tools you use, do any of them have a recommendation functionality, either main, or an additional functionality? (Ex: "Similar papers" sections in libraries, Semantic Scholar feeds)
- If so, do you feel that the papers you come into contact with through these tools/functionalities are different from those you find through other tools? How?

- Do you use any tools that have functionalities based on LLMs? (Ex: ChatGPT, Elicit, ResearchRabbit, Scholarcy)
- What do you usually do with these LLMs tools?
- On the issue specifically of finding papers, have these LLM-based tools played any role?
- If so, do you feel that the papers you come into contact with through these tools/functionalities are different from those you find through other tools? How?

- Would you say these tools and practices vary according to the stage of work?
- If you work together with other people at any stage of your work, would you say it also varies with that?

- Do you have any work/paper that you have done, perhaps more recently, in which you can recount your process of searching for papers more directly?
- How did you find the most interesting papers that were used in this work?

- Do you use any tool or tools to save papers?
- How do you organize the papers within these tools?
- In this organization, what would you say is the relationship between the papers that are stored in the same group/tag/folder?
- Why did you decide to organize it this way?

- Besides this tool for saving, do you use any other tool for organizing papers?
- Do you make any kind of "index cards" (or annotation, record) of the papers at any point in your work?

- How would you say these tools and practices vary according to the stage of work? (If the answer is more related to systematic reviews, ask more about other cases)

- What do you understand by "similarity" when we are talking about papers?
- What do you understand by "connection" when we are talking about papers?
- Do you think it is distinct from the idea of similarity?
- When you are reading an paper, what aspects suggest to you that it is similar to another?
- When you are reading an paper, what aspects suggest to you that it is connected to another? (It is possible that the answer is very related to citations, instigate from there)

- In the organization you described, are the connections between papers stored in any way?
- Are "connected" papers, in the way you described, stored together?
- What kind of information would you say you store about papers in your organization?
- Do you have any organization for notes from your research work?
- If so, is it connected to your organization for papers?

B.1 Study 1 Participant Quotes

Translated from Portuguese

Q1 "Since I don't read prospectively that much... like, 'Ah, I'll read this article for fun.' It doesn't end up being very useful for me." (P9)

Q2 "Based on what you're seeing, do [tools] say 'I recommend these articles'? Look, if that exists, I don't know about it. If that exists and works well, I don't know about it." (P5)

Q3 "In the end, it must have a black box... it does its queries there, but I don't think it's enough for me to tell you, 'wow, Research Rabbit, wow, it changed my life.'" (P2)

Q4 "If I've already gotten that far, if I've already gone through the trouble of finding that paper, it's often very difficult to find other papers afterwards. So that's where the recommendation is useful for me, you know?" (P2)

Q5 "Sometimes I'm in an article here and... oh, there's also this here and so on. I end up clicking... It seems they are usually the articles I didn't find through the search." (P10)

Q6 "That helps in a way you can't imagine... see if there's an image similar to what you're thinking, because that image probably came from an article that interests you." (P8)

Q7 "So I tried to use it like, give me references... that talk about the use of authoring tools by teachers... Then it gave me, it made it up." (P3)

Q8 "We used Bard to look for articles about Bard. And it returned a reference to us, but this reference didn't exist." (P10)

Q9 "I trust my own criteria more, I don't really know what criteria it's using." (P8)

Q10 "Yeah, for brainstorming, and stuff, right? Since it hallucinates, it gives you tips... sometimes it sends you to places you hadn't thought of, gives you different ideas." (P9)

Q11 "I take this citation, whether it's right or not, I throw it into... Google Scholar, and then it finds something very close by the keywords that it... managed to generate." (P7)

Q12 "then I went to Scopus, I started with queries... and from there a list comes, right, and then I was filtering, okay... in Scopus you can, what are the groups from Brazil, then you can sort by citation and such." (P2)

Q13 "The way I use it, Scopus is a more structured search, I make a formalized search string and in Scholar... it's a freer environment, you know? So I write a more generic term and I go finding things." (P10)

Q14 "I think the more on the edge you are, the more important it is for you to move away from... looking for a specific paper and much more to understand the ecosystem, to understand the group." (P2)

Q15 "The first time I looked at it and said, how cool, it shows who cited whom... Man, for me that's a gimmick. [...] This bunch of papers cited this one. Okay, that could be for N reasons." (P2)

Q16 "So the article itself is a search tool." (P8)

Q17 "I read on the tablet, I mark things, and then I write down... questions and opportunities, right?" (P10)

Q18 "Yeah, a reminder, almost like a scrapbook, like, ah, maybe I'll need this tomorrow, then I write it down." (P9)

Q19 "also a bit of a Wiki vibe, which is like, you know, you have a subject and then you describe that fellow a bit... I can unbelievably quickly transmit a concept, transmit a thing." (P1)

Q20 "I'm taking several notes on methods and such, I have the references. And then, when I'm going to write a paper... I go to these notes and see what I wrote, what can go into the paper." (P3)

Q21 "We build the article. I like to already put it in a template, structure... the sections and such... and then already have a bib file there to save the references and build it that way." (P4)

Q22 "we have a habit of always making partial presentations... So, like, these partial presentations end up... making our curation process happen, you know?" (P6)

Q23 "I started to notice several, quote-unquote, 'useless' papers, and parts of them are very useful... the story the guy is telling... is not 100% useful to me. But some experiment he did in between is very useful." (P2)

Q24 "So when you read, you think about that, right? How am I going to approach the reading? What part can you ignore in the article, for example? I always think about that." (P9)

Q25 "Two articles can be connected either because one references the other... in the sense of 'I am continuing that work'" (P1)

Q26 "There is a stronger type of connection, which are articles that... are presenting a line of work from a set of authors." (P9)

Q27 "it's simple to find the similarity, because, basically, it's the same literature, it's the same problem, only the method is different." (P4)

Q28 "Similarity. I think it depends on the researcher's point of view." (P7)

Q29 "In this organization, they will be stored together. But the connection is not explicit in this storage." (P4)

Q30 "I had already done this for a previous paper... But, sometimes, I don't have much... Yeah, patience or discipline." (P4)

Q31 "Understanding what the queries were, what the items were, helps me understand more of my field." (P2)

Q32 "It has been useful and different in the sense of assistance. So there's this thing of kind of talking to an intern, you know? It helps me understand things that I'm already trying to understand." (P10)

C
User Study Script

Translated from Portuguese

Session 1 - Initial 15–20 minute conversation

Quick explanation of what the tool is.

If you participated in the first study:

- We wanted to first check if anything has changed regarding what you told us in the first study:
  - Your research work
  - The tools you use for organizing papers
  - Your main research topic
- Can you list some subtopics within this domain with which you have a more direct connection?
- Thinking about these subtopics, can you recall any recent research (formal or informal) you have done on one of them?

The proposal for the next stage of this study is to select three of these subtopics and create collections with example texts. Ideally, two of these collections would be on topics with which you have more experience, and you could provide us with a few (3-4+) texts to use. The third would be a topic that you are familiar with, can point to at least one reference text for, but haven't had as much direct contact with. If you prefer, we can have more examples for each. If you already have collections in Zotero or another reference manager, we can start from those.

- For each subtopic as we assemble the collections:
  - Can you explain it to me briefly?
  - Is this a topic you have researched in more depth, or something you have less contact with?
  - Can you remember some papers that would be directly related to this subtopic? Can you show me or pass me the examples directly?

That is all we need for the first session. In addition to what we discussed, for the two collections you are more familiar with, it would be great if you could bring one or more texts to our next conversation that you would include, besides the ones you already mentioned. You can send me these texts before our next conversation if you prefer.

Session 2 - Think-aloud, 3 phases (1h)

The objective of this second phase is to conduct a "think-aloud" style test, which is a format where you will interact with the tool naturally, but narrate your thoughts out loud as you perform each action. This can include doubts, impressions, decisions, or anything else that comes to mind - the important thing is to share your reasoning as it happens. I will guide you lightly throughout the test, and at times ask you to perform some specific actions. If you prefer to follow a different path as we go, you can let me know.

1. Review and initial visualization

In Zotero, we organized three collections that reflect the subtopics you brought up in the first conversation, using the texts you provided. When you access one of them, our tool will present initial suggestions in the sidebar, generated from the papers in the collection. As you interact with these suggestions, new ones will be proposed based on your actions. To begin, you can choose any of the collections.

- **About the collection**
  - What is this set of texts about?
  - Let's review how you structured this collection. What would you say are the themes, projects, or other criteria that connect these papers? What defines the inclusion of a paper here?
- **About the suggestions**
  - Please consider the suggestions that were generated by the tool. You can interact with them freely, but I would like to ask you to look at the complete set first.
  - Consider the first set of suggestions presented by the tool.
    - What connections or themes do you think are being prioritized in these suggestions?
    - In what way, if any, would you say these suggestions relate to the original intent of this collection of texts?

  - Please choose one suggestion to explore a little further, whichever one you prefer.
    - What connections or themes do you think are being prioritized in this specific suggestion?
    - In what way, if any, would you say this suggestion relates to the original intent of this collection of texts?
    - Why did you choose this suggestion for us to analyze?

- **About the linked text passages**
  - Please consider the passages linked within the suggestions. You can interact with them freely.
    - (When interacting with a passage) In what way, if any, would you say this passage relates to the suggestion, and to the collection of papers?
    - And between this passage and your own perception of the paper?
    - Returning to the suggestion: any observations about the other passages that were suggested along with it?
    - Why did you choose these passages for us to analyze?

You can look at as many other suggestions and passages as you wish.

(We will repeat this for the other two collections, up to the point of discussing the text passages.)

Now, I would like you to please choose one of these collections and think of a text you would like to add that is not yet in it. It could be a text you already know and consider relevant, or one you want to test now. While we are doing the next part, your text will be processing.

(To facilitate interaction, given the limitations of Zotero Web, an upload button has been added to the tool)

2. Directed interaction with collection content

Now, the idea is to choose some collections to interact with more deeply. I will ask you to choose a collection to add a note to, and at least one text to highlight some passages in. They can be from the same collection, or not.

- **Text highlighting and generation**
  - Please, I would like you to choose some related passages from the selected text to highlight. If you are used to doing this, you can do it in exactly the same way you normally do. You can choose the

    same text you just read, or you can navigate to another one if you prefer.
  - Examine the new suggestion generated by the tool from the content you highlighted. Interact with it as you wish. In what way, if any, would you say this suggestion relates to the original intent of this collection of texts?

- **Note addition and generation**

  - Please add a note to one of the collections.
  - Examine the new suggestion generated by the tool from the note you added. Interact with it as you wish. In what way, if any, would you say this suggestion relates to the original intent of this collection of texts?

- **Generation from an added text**

  - Let's add a new paper to the collection.

    * What motivated the choice of this paper? What relationship do you see between it and the other texts that make up this collection?
    * After the addition, the tool generates a new suggestion based on the paper's content. Feel free to interact with it. Do you feel that this suggestion relates to the purpose of the collection or to the motivation for your choice?

- Would you like to repeat any of these actions in another one of the collections?

3. Reflective synthesis / review of general points

- During navigation and interaction, did you generally feel that the tool focused on the passages of the papers that you would also focus on, considering the context of each collection?

  - If not, would you say they were positive or negative divergences?

- Overall, how would you say the suggestions presented seemed to align with what you had in mind for each collection?
- And regarding the connections between papers proposed by the tool within the suggestions - did they make sense to you? Would you generally denote them in a similar way?

- Have you used LLM-based tools with a similar character to this one? What was your experience like?
- Have you used recommendation tools with a similar character to this one? What was your experience like?
- Based on your experience and work process, would you say that this tool could fit well, or better, at any particular moment in your process?

Thank you, that's everything we needed. We will send you all the content of the suggestions that were generated from your texts in an email, in case they might be useful to you in some way.

C.1 Participant Quotes (Second Study)

Translated from Portuguese

Q1 "The general theme is great, the problem is the citation." (P4)

Q2 "The what I would do is, like, ah okay, you're describing a situation, let me try to find the key terms you used in this description here, do the Ctrl+F direct." (P4)

Q3 "bad, it's just a title." (P3)

Q4 "The abstract is a first approach, but perhaps with information that is not so detailed." (P5)

Q5 "very complicated to draw that conclusion from that snippet of an equation." (P1)

Q6 "It's like telling me it uses an A100 GPU, but not telling me anything else. It's just that, isolated and useless." (P7)

Q7 "I don't imagine it would have the capacity to understand that here we have these arrows, that here you have this big circle... This, for me, is an interpretive dimension that I suppose language models haven't well developed yet." (P6)

Q8 "here for me you actually lose a small semantic layer, because the italicization here was for emphasis" (P5)

Q9 "it seems to me that it hallucinated something." (P4)

Q10 "No, in fact it's the other way around. We use the experimental results to validate the numerical ones. We create the numerical model based on the experimental one, because the experimental is the real one." (P1)

Q11 "Of course. Because they are giving me a summary of what is being discussed, so I can get a summarized view here without having to open the whole text." (P7)

Q12 "It failed miserably at summarizing the papers. It told me nothing I couldn't have guessed from the title." (P7)

Q13 "I wanted... if it could be a little, not super verbose, you know, but a little more informative." (P3)

Q14 "No, so, here I think it's the same mistake as before. For example, it said that all texts talk about stainless steel tubular connections, and no, not all of them do. Same thing here. It says, all texts use finite element analysis. No, not all of them do." (P1)

Q15 "I found it a bit generic because it's 'iterative processes to refine segmentations and optimize models'. Like, this thing about iterating... Any neural network is trained that way." (P4)

Q16 "Of course. Because they are giving me a summary of what is being discussed, so I can get a summarized view here without having to open the whole text." (P1)

Q17 "I think that to make a, let's say, more initial thing, I think it makes, for me, it made more sense... it gives insights or even a direction for reading." (P3)

Q18 "It seems like a way for me to navigate internally in the text by going to this theme here... I navigated more quickly." (P6)

Q19 "It's very automatic for me the desire to navigate with the Ctrl+F." (P6)

Q20 "Oh, for sure. For example, when it talked about parametric analyses... It is in fact something that was common in all the texts." (P1)

Q21 "So, in this case, it correctly identified that this is a theme that appears only in these two and that it is a theme that exists." (P4)

Q22 "It's because here it is taking specific aspects. It is slicing the theme into a topic. And for each method, it is trying to see how the method aligns with this topic." (P5)

Q23 "If the developers realize that, look, we have to help people's process instead of doing the process for them, then maybe we'll arrive at a very interesting key to reading that will help real scientific work happen. Otherwise, we'll continue creating AI Slop." (P6)

Q24 "It was very reminiscent of the work I did when I wrote my dissertation, where you have here big themes that possibly repeat inside these texts and how they are potentially crossing over." (P6)

Q25 "This first one got the core of the matter... For me, this here was super 100% aligned." (P4)

Q26 "It happened what I... not what I wanted, but what I imagined would happen, and that was actually what I wanted to happen for me. The subjects were well segregated." (P8)

Q27 "Yes, because not where I would focus, but where I might want to focus." (P5)

Q28 "I didn't even know what that term was. But it's a key term that you as a reader notice. Wait, this is important. So let's fixate on this term here. And even look for other articles that use this term and use it as a basis to better learn the niche." (P4)

Q29 "No, in fact it's the other way around. We use the experimental results to validate the numerical ones. We create the numerical model based on the experimental one, because the experimental is the real one." (P1)

Q30 "'Several texts emphasize the importance of visual and structured data representations...' This text here, which is in this reference I'm looking at now, says nothing about visuality." (P6)

Q31 "He was deceived in a very easy way here on this subject, basically." (P8)

Q32 "This is consistent. It's exactly that, from the texts I sent... It is really something that is common in all the texts." (P1)

Q33 "My idea for the note was to ask a question to the documents... to see if I could find something new." (P6)

Q34 "I think it's very useful for... bootstrapping the thing. You look at it and you already know where to go." (P8)

Q35 "I marked this one and it expanded a bit on what I had marked. I think this kind of thing is useful." (P3)

Q36 "Oh, that's interesting. It's an interesting aspect that I hadn't thought about, of the dataset being a central point. That's a good point, I hadn't thought about that." (P5)

Q37 "What I had asked for in the note was like, I wonder if others use it? So, I expected a bridge to another one. And it didn't make a bridge." (P4)

Q38 "I think the problem here is that the word 'random' appears. And then I think he gets lost... I think he saw the word 'random' and said 'ah, he wants to talk about randomness', but the context is another." (P7)

Q39 "The extension here seems to have read a linguistic resource as being the substance of what the text was saying... The machine is reading and synthesizing something that was made for us, to maintain the flow of reasoning." (P6)

Q40 "I hated this one, because what it gave me about this article was something about 'Prior Work'. It took a sentence that talks about previous works that doesn't tell me much about this specific work." (P7)

Q41 "It doesn't tell me anything about the article. It's like telling me it uses an A100 GPU, but not telling me anything else. It's just that, isolated and useless." (P7)

Q42 "That when it can't fit [a paper] into the narrative, it excludes it." (P7)

Q43 "if it marked that thing that had nothing to do with it... it increases my distrust even of what seemed useful." (P6)

Q44 "vision of literature, related works, sometimes is not so interesting, what you want is the method so you can compare the two methods." (P5)

Q45 "These articles present the procedures. They expose the specific boundary conditions for the study, show the procedures adopted in the tests, how the modeling was developed and, of course, the results." (P1)

Q46 "In this set of papers, you're wanting to focus on this rigor, on the proofs, on the theorems, not so much on the method." (P5)

Q47 "What guarantees me that somewhere else in the rest of this text, if I haven't actually read it, there won't be another piece of information that's the inverse of this one, or something more interesting that isn't marked?" (P6)

Q48 "It doesn't show how the papers relate to each other... It doesn't say that these two works are very related... maybe one is a continuation of the other." (P7)

Q49 "I didn't even know what that term was. But it's a key term that you as a reader notice. Wait, this is important... and use it as a basis to better learn the niche." (P4)

Q50 "I don't know if I would do something very exploratory. Because what I read here, I read some themes that I can't say what is important, what is not, what is for me, what is not." (P8)

Q51 "this is a quick way to find the detail, maybe even copy the own text... and use it in your 'related work' section." (P4)

Q52 "when you are doing a systematic literature review... it gives you these insights... or even a direction for reading." (P3)

Q53 "I would say it would be really cool if I had a body of text that I already know, and I could keep dropping new things in and it will make links between them." (P8)

Q54 "Usually people do a literature review in two moments: either before starting the work, or when they are almost submitting the article and want to fit that article into the middle of the literature. I think that in either of those two moments, this can be useful." (P7)

Q55 "If the developers realize that, look, we have to help people's process instead of doing the process for them, then maybe we'll arrive at a very interesting key to reading that will help real scientific work happen. Otherwise, we'll continue creating AI Slop, scientific mush." (P6)

Q56 "It feels like I'm outsourcing my reading, in the end... a single error had a cascading effect... to increase my distrust even of what seemed useful." (P6)

Q57 "Especially if you can establish your context for the tool. What you want to look at in the papers." (P5)

Q58 "My intention, when I grouped them, if I wanted to recall this information, would be to have an overview of each one and have a chronology to understand where one builds on the other." (P4)

Q59 "It doesn't show how the papers relate to each other... It doesn't say that these two works are very related... maybe one is a continuation of the other." (P7)

Q60 "Again, if it were a 'deep dive', like ChatGPT's 'deep research' mode, it would be interesting. To give it several texts and have it give me an extract, almost like doing a meta-analysis. That could be something to have." (P8)